NONPARAMETRIC ESTIMATION OF AVERAGE TREATMENT EFFECTS UNDER EXOGENEITY: A REVIEW*

Guido W. Imbens

Abstract—Recently there has been a surge in econometric work focusing on estimating average treatment effects under various sets of assumptions. One strand of this literature has developed methods for estimating average treatment effects for a binary treatment under assumptions variously described as exogeneity, unconfoundedness, or selection on observables. The implication of these assumptions is that systematic (for example, average or distributional) differences in outcomes between treated and control units with the same values for the covariates are attributable to the treatment. Recent analysis has considered estimation and inference for average treatment effects under weaker assumptions than typical of the earlier literature by avoiding distributional and functional-form assumptions. Various methods of semiparametric estimation have been proposed, including estimating the unknown regression functions, matching, methods using the propensity score such as weighting and blocking, and combinations of these approaches. In this paper I review the state of this literature and discuss some of its unanswered questions, focusing in particular on the practical implementation of these methods, the plausibility of this exogeneity assumption in economic applications, the relative performance of the various semiparametric estimators when the key assumptions (unconfoundedness and overlap) are satisfied, alternative estimands such as quantile treatment effects, and alternate methods such as Bayesian inference.

I. Introduction

SINCE the work by Ashenfelter (1978), Card and Sullivan (1988), Heckman and Robb (1984), Lalonde (1986), and others, there has been much interest in econometric methods for estimating the effects of active labor market programs such as job search assistance or classroom teaching programs. This interest has led to a surge in theoretical work focusing on estimating average treatment effects under various sets of assumptions. See for general surveys of this literature Angrist and Krueger (2000), Heckman, LaLonde, and Smith (2000), and Blundell and Costa-Dias (2002).

One strand of this literature has developed methods for estimating the average effect of receiving or not receiving a binary treatment under the assumption that the treatment satisfies some form of exogeneity. Different versions of this assumption are referred to as unconfoundedness (Rosenbaum & Rubin, 1983a), selection on observables (Barnow, Cain, & Goldberger, 1980; Fitzgerald, Gottschalk, & Moffitt, 1998), or conditional independence (Lechner, 1999). In the remainder of this paper I will use the terms unconfoundedness and exogeneity interchangeably to denote the assumption that the receipt of treatment is independent of the potential outcomes with and without treatment if certain observable covariates are held constant. The implication of these assumptions is that systematic (for example, average or distributional) differences in outcomes between treated and control units with the same values for these covariates are attributable to the treatment.

Much of the recent work, building on the statistical literature by Cochran (1968), Cochran and Rubin (1973), Rubin (1973a, 1973b, 1977, 1978), Rosenbaum and Rubin (1983a, 1983b, 1984), Holland (1986), and others, considers estimation and inference without distributional and functional form assumptions. Hahn (1998) derived efficiency bounds assuming only unconfoundedness and some regularity conditions and proposed an efficient estimator. Various alternative estimators have been proposed given these conditions. These estimation methods can be grouped into five categories: (i) methods based on estimating the unknown regression functions of the outcome on the covariates (Hahn, 1998; Heckman, Ichimura, & Todd, 1997, 1998; Imbens, Newey, & Ridder, 2003), (ii) matching on covariates (Rosenbaum, 1995; Abadie and Imbens, 2002) (iii) methods based on the propensity score, including blocking (Rosenbaum & Rubin, 1984) and weighting (Hirano, Imbens, & Ridder, 2003), (iv) combinations of these approaches, for example, weighting and regression (Robins & Rotnizky, 1995) or matching and regression (Abadie & Imbens, 2002), and (v) Bayesian methods, which have found relatively little following since Rubin (1978). In this paper I will review the state of this literature—with particular emphasis on implications for empirical work—and discuss some of the remaining questions.

The organization of the paper is as follows. In section II I will introduce the notation and the assumptions used for identification. I will also discuss the difference between population- and sample-average treatment effects. The recent econometric literature has largely focused on estimation of the population-average treatment effect and its counterpart for the subpopulation of treated units. An alternative, following the early experimental literature (Fisher, 1925; Neyman, 1923), is to consider estimation of the average effect of the treatment for the units in the sample. Many of the estimators proposed can be interpreted as estimating either the average treatment effect for the sample at hand, or the average treatment effect for the population. Although the choice of estimand may not affect the form of the estimator, it has implications for the efficiency bounds and for the form of estimators of the asymptotic variance; the variance of estimators for the sample average treatment effect are

Received for publication October 22, 2002. Revision accepted for publication June 4, 2003.

* University of California at Berkeley and NBER

This paper was presented as an invited lecture at the Australian and European meetings of the Econometric Society in July and August 2003. I am also grateful to Joshua Angrist, Jane Herr, Caroline Hoxby, Charles Manski, Xiangyi Meng, Robert Moffitt, and Barbara Sianesi, and two referees for comments, and to a number of collaborators, Alberto Abadie, Joshua Angrist, Susan Athey, Gary Chamberlain, Keisuke Hirano, V. Joseph Hotz, Charles Manski, Oscar Mitnik, Julie Mortimer, Jack Porter, Whitney Newey, Geert Ridder, Paul Rosenbaum, and Donald Rubin for many discussions on the topics of this paper. Financial support for this research was generously provided through NSF grants SBR 9818644 and SES 0136789 and the Giannini Foundation.

generally smaller. In section II, I will also discuss alternative estimands. Almost the entire literature has focused on average effects. However, in many cases such measures may mask important distributional changes. These can be captured more easily by focusing on quantiles of the distributions of potential outcomes, in the presence and absence of the treatment (Lehman, 1974; Docksum, 1974; Firpo, 2003).

In section III, I will discuss in more detail some of the recently proposed semiparametric estimators for the average treatment effect, including those based on regression, matching, and the propensity score. I will focus particularly on implementation, and compare the different decisions faced regarding smoothing parameters using the various estimators.

In section IV, I will discuss estimation of the variances of these average treatment effect estimators. For most of the estimators introduced in the recent literature, corresponding estimators for the variance have also been proposed, typically requiring additional nonparametric regression. In practice, however, researchers often rely on bootstrapping, although this method has not been formally justified. In addition, if one is interested in the average treatment effect for the sample, bootstrapping is clearly inappropriate. Here I discuss in more detail a simple estimator for the variance for matching estimators, developed by Abadie and Imbens (2002), that does not require additional nonparametric estimation.

Section V discusses different approaches to assessing the plausibility of the two key assumptions: exogeneity or unconfoundedness, and overlap in the covariate distributions. The first of these assumptions is in principle untestable. Nevertheless a number of approaches have been proposed that are useful for addressing its credibility (Heckman and Hotz, 1989; Rosenbaum, 1984b). One may also wish to assess the responsiveness of the results to this assumption using a sensitivity analysis (Rosenbaum & Rubin, 1983b; Imbens, 2003), or, in its extreme form, a bounds analysis (Manski, 1990, 2003). The second assumption is that there exists appropriate overlap in the covariate distributions of the treated and control units. That is effectively an assumption on the joint distribution of observable variables. However, as it only involves inequality restrictions, there are no direct tests of this null. Nevertheless, in practice it is often very important to assess whether there is sufficient overlap to draw credible inferences. Lacking overlap for the full sample, one may wish to limit inferences to the average effect for the subset of the covariate space where there exists overlap between the treated and control observations.

In Section VI, I discuss a number of implementations of average treatment effect estimators. The first set of implementations involve comparisons of the nonexperimental estimators to results based on randomized experiments, allowing direct tests of the unconfoundedness assumption. The second set consists of simulation studies—using data created either to fulfill the unconfoundedness assumption or to fail it a known way—designed to compare the applicability of the various treatment effect estimators in these diverse settings.

This survey will not address alternatives for estimating average treatment effects that do not rely on exogeneity assumptions. This includes approaches where selected observed covariates are not adjusted for, such as instrumental variables analyses (Björklund & Moffit, 1987; Heckman & Robb, 1984; Imbens & Angrist, 1994; Angrist, Imbens, & Rubin, 1996; Ichimura & Taber, 2000; Abadie, 2003a; Chernozhukov & Hansen, 2001). I will also not discuss methods exploiting the presence of additional data, such as difference in differences in repeated cross sections (Abadie, 2003b; Blundell et al., 2002; Athey and Imbens, 2002) and regression discontinuity where the overlap assumption is violated (van der Klaauw, 2002; Hahn, Todd, & van der Klaauw, 2000; Angrist & Lavy, 1999; Black, 1999; Lee, 2001; Porter, 2003). I will also limit the discussion to binary treatments, excluding models with static multivalued treatments as in Imbens (2000) and Lechner (2001) and models with dynamic treatment regimes as in Ham and LaLonde (1996), Gill and Robins (2001), and Abbring and van den Berg (2003). Reviews of many of these methods can be found in Shadish, Campbell, and Cook (2002), Angrist and Krueger (2000), Heckman, LaLonde, and Smith (2000), and Blundell and Costa-Dias (2002).

II. Estimands, Identification, and Efficiency Bounds

A. *Definitions*

In this paper I will use the potential-outcome notation that dates back to the analysis of randomized experiments by Fisher (1935) and Neyman (1923). After being forcefully advocated in a series of papers by Rubin (1974, 1977, 1978), this notation is now standard in the literature on both experimental and nonexperimental program evaluation.

We begin with N units, indexed by $i = 1, \ldots, N$, viewed as drawn randomly from a large population. Each unit is characterized by a pair of potential outcomes, $Y_i(0)$ for the outcome under the control treatment and $Y_i(1)$ for the outcome under the active treatment. In addition, each unit has a vector of characteristics, referred to as covariates, pretreatment variables, or exogenous variables, and denoted by X_i.[1] It is important that these variables are not affected by the treatment. Often they take their values prior to the unit being exposed to the treatment, although this is not sufficient for the conditions they need to satisfy. Importantly, this vector of covariates can include lagged outcomes.

[1] Calling such variables exogenous is somewhat at odds with several formal definitions of exogeneity (e.g., Engle, Hendry, & Richard, 1974), as knowledge of their distribution can be informative about the average treatment effects. It does, however, agree with common usage. See for example, Manski et al. (1992, p. 28). See also Frölich (2002) and Hirano et al. (2003) for additional discussion.

Finally, each unit is exposed to a single treatment; $W_i = 0$ if unit i receives the control treatment, and $W_i = 1$ if unit i receives the active treatment. We therefore observe for each unit the triple (W_i, Y_i, X_i), where Y_i is the realized outcome:

$$Y_i \equiv Y_i(W_i) = \begin{cases} Y_i(0) & \text{if } W_i = 0, \\ Y_i(1) & \text{if } W_i = 1. \end{cases}$$

Distributions of (W, Y, X) refer to the distribution induced by the random sampling from the superpopulation.

Several additional pieces of notation will be useful in the remainder of the paper. First, the propensity score (Rosenbaum and Rubin, 1983a) is defined as the conditional probability of receiving the treatment,

$$e(x) \equiv \Pr(W = 1|X = x) = \mathbb{E}[W|X = x].$$

Also, define, for $w \in \{0, 1\}$, the two conditional regression and variance functions

$$\mu_w(x) \equiv \mathbb{E}[Y(w)|X = x], \quad \sigma_w^2(x) \equiv \mathbb{V}(Y(w)|X = x).$$

Finally, let $\rho(x)$ be the conditional correlation coefficient of $Y(0)$ and $Y(1)$ given $X = x$. As one never observes $Y_i(0)$ and $Y_i(1)$ for the same unit i, the data only contain indirect and very limited information about this correlation coefficient.[2]

B. *Estimands: Average Treatment Effects*

In this discussion I will primarily focus on a number of average treatment effects (ATEs). This is less limiting than it may seem, however, as it includes averages of arbitrary transformations of the original outcomes. Later I will return briefly to alternative estimands that cannot be written in this form.

The first estimand, and the most commonly studied in the econometric literature, is the population-average treatment effect (PATE):

$$\tau^P = \mathbb{E}[Y(1) - Y(0)].$$

Alternatively we may be interested in the population-average treatment effect for the treated (PATT; for example, Rubin, 1977; Heckman & Robb, 1984):

$$\tau_T^P = \mathbb{E}[Y(1) - Y(0)|W = 1].$$

Heckman and Robb (1984) and Heckman, Ichimura, and Todd (1997) argue that the subpopulation of treated units is often of more interest than the overall population in the context of narrowly targeted programs. For example, if a program is specifically directed at individuals disadvantaged in the labor market, there is often little interest in the effect of such a program on individuals with strong labor market attachment.

I will also look at sample-average versions of these two population measures. These estimands focus on the average of the treatment effect in the specific sample, rather than in the population at large. They include the sample-average treatment effect (SATE)

$$\tau^S = \frac{1}{N} \sum_{i=1}^{N} [Y_i(1) - Y_i(0)],$$

and the sample-average treatment effect for the treated (SATT)

$$\tau_T^S = \frac{1}{N_T} \sum_{i:W_i=1} [Y_i(1) - Y_i(0)],$$

where $N_T = \sum_{i=1}^{N} W_i$ is the number of treated units. The SATE and the SATT have received little attention in the recent econometric literature, although the SATE has a long tradition in the analysis of randomized experiments (for example, Neyman, 1923). Without further assumptions, the sample contains no information about the PATE beyond the SATE. To see this, consider the case where we observe the sample $(Y_i(0), Y_i(1), W_i, X_i)$, $i = 1, \ldots, N$; that is, we observe both potential outcomes for each unit. In that case $\tau^S = \sum_i [Y_i(1) - Y_i(0)]/N$ can be estimated without error. Obviously, the best estimator for the population-average effect τ^P is τ^S. However, we cannot estimate τ^P without error even with a sample where all potential outcomes are observed, because we lack the potential outcomes for those population members not included in the sample. This simple argument has two implications. First, one can estimate the SATE at least as accurately as the PATE, and typically more so. In fact, the difference between the two variances is the variance of the treatment effect, which is zero only when the treatment effect is constant. Second, a good estimator for one ATE is automatically a good estimator for the other. One can therefore interpret many of the estimators for PATE or PATT as estimators for SATE or SATT, with lower implied standard errors, as discussed in more detail in section IIE.

A third pair of estimands combines features of the other two. These estimands, introduced by Abadie and Imbens (2002), focus on the ATE conditional on the sample distribution of the covariates. Formally, the conditional ATE (CATE) is defined as

$$\overline{\tau(X)} = \frac{1}{N} \sum_{i=1}^{N} \mathbb{E}[Y_i(1) - Y_i(0)|X_i],$$

and the SATE for the treated (CATT) is defined as

$$\overline{\tau(X)_T} = \frac{1}{N_T} \sum_{i:W_i=1} \mathbb{E}[Y_i(1) - Y_i(0)|X_i].$$

[2] As Heckman, Smith, and Clemens (1997) point out, however, one can draw some limited inferences about the correlation coefficient from the shape of the two marginal distributions of $Y(0)$ and $Y(1)$.

Using the same argument as in the previous paragraph, it can be shown that one can estimate CATE and CATT more accurately than PATE and PATT, but generally less accurately than SATE and SATT.

The difference in asymptotic variances forces the researcher to take a stance on what the quantity of interest is. For example, in a specific application one can legitimately reach the conclusion that there is no evidence, at the 95% level, that the PATE is different from zero, whereas there may be compelling evidence that the SATE and CATE are positive. Typically researchers in econometrics have focused on the PATE, but one can argue that it is of interest, when one cannot ascertain the sign of the population-level effect, to know whether one can determine the sign of the effect for the sample. Especially in cases, which are all too common, where it is not clear whether the sample is representative of the population of interest, results for the sample at hand may be of considerable interest.

C. *Identification*

We make the following key assumption about the treatment assignment:

ASSUMPTION 2.1 (UNCONFOUNDEDNESS):

$$(Y(0),\ Y(1)) \perp W|X.$$

This assumption was first articulated in this form by Rosenbaum and Rubin (1983a), who refer to it as "ignorable treatment assignment." Lechner (1999, 2002) refers to this as the "conditional independence assumption." Following work by Barnow, Cain, and Goldberger (1980) in a regression setting it is also referred to as "selection on observables."

To see the link with standard exogeneity assumptions, suppose that the treatment effect is constant: $\tau = Y_i(1) - Y_i(0)$ for all i. Suppose also that the control outcome is linear in X_i:

$$Y_i(0) = \alpha + X_i'\beta + \varepsilon_i,$$

with $\varepsilon_i \perp X_i$. Then we can write

$$Y_i = \alpha + \tau \cdot W_i + X_i'\beta + \varepsilon_i.$$

Given the assumption of constant treatment effect, unconfoundedness is equivalent to independence of W_i and ε_i conditional on X_i, which would also capture the idea that W_i is exogenous. Without this assumption, however, unconfoundedness does not imply a linear relation with (mean-)independent errors.

Next, we make a second assumption regarding the joint distribution of treatments and covariates:

ASSUMPTION 2.2 (OVERLAP):

$$0 < \Pr(W = 1|X) < 1.$$

For many of the formal results one will also need smoothness assumptions on the conditional regression functions and the propensity score [$\mu_w(x)$ and $e(x)$], and moment conditions on $Y(w)$. I will not discuss these regularity conditions here. Details can be found in the references for the specific estimators given below.

There has been some controversy about the plausibility of Assumptions 2.1 and 2.2 in economic settings, and thus about the relevance of the econometric literature that focuses on estimation and inference under these conditions for empirical work. In this debate it has been argued that agents' optimizing behavior precludes their choices being independent of the potential outcomes, whether or not conditional on covariates. This seems an unduly narrow view. In response I will offer three arguments for considering these assumptions.

The first is a statistical, data-descriptive motivation. A natural starting point in the evaluation of any program is a comparison of average outcomes for treated and control units. A logical next step is to adjust any difference in average outcomes for differences in exogenous background characteristics (exogenous in the sense of not being affected by the treatment). Such an analysis may not lead to the final word on the efficacy of the treatment, but its absence would seem difficult to rationalize in a serious attempt to understand the evidence regarding the effect of the treatment.

A second argument is that almost any evaluation of a treatment involves comparisons of units who received the treatment with units who did not. The question is typically not whether such a comparison should be made, but rather which units should be compared, that is, which units best represent the treated units had they not been treated. Economic theory can help in classifying variables into those that need to be adjusted for versus those that do not, on the basis of their role in the decision process (for example, whether they enter the utility function or the constraints). Given that, the unconfoundedness assumption merely asserts that all variables that need to be adjusted for are observed by the researcher. This is an empirical question, and not one that should be controversial as a general principle. It is clear that settings where some of these covariates are not observed will require strong assumptions to allow for identification. Such assumptions include instrumental variables settings where some covariates are assumed to be independent of the potential outcomes. Absent those assumptions, typically only bounds can be identified (as in Manski, 1990, 2003).

A third, related argument is that even when agents choose their treatment optimally, two agents with the same values for observed characteristics may differ in their treatment choices without invalidating the unconfoundedness assumption if the difference in their choices is driven by differences in unobserved characteristics that are themselves unrelated to the outcomes of interest. The plausibility of this will depend critically on the exact nature of the optimization

process faced by the agents. In particular it may be important that the objective of the decision maker is distinct from the outcome that is of interest to the evaluator. For example, suppose we are interested in estimating the average effect of a binary input (such as a new technology) on a firm's output.[3] Assume production is a stochastic function of this input because other inputs (such as weather) are not under the firm's control: $Y_i = g(W, \varepsilon_i)$. Suppose that profits are output minus costs ($\pi_i = Y_i - c_i \cdot W_i$), and also that a firm chooses a production level to maximize expected profits, equal to output minus costs, conditional on the cost of adopting new technology,

$$W_i = \arg\max_{w \in \{0,1\}} \mathbb{E}[\pi(w)|c_i]$$

$$= \arg\max_{w \in \{0,1\}} \mathbb{E}[g(w, \varepsilon_i) - c_i \cdot w|c_i],$$

implying

$$W_i = 1\{\mathbb{E}[g(1, \varepsilon) - g(0, \varepsilon_i) \geq c_i|c_i]\} = h(c_i).$$

If unobserved marginal costs c_i differ between firms, and these marginal costs are independent of the errors ε_i in the firms' forecast of production given inputs, then unconfoundedness will hold, as

$$(g(0, \varepsilon), g(1, \varepsilon_i)) \perp c_i.$$

Note that under the same assumptions one cannot necessarily identify the effect of the input on profits, for $(\pi_i(0), \pi(1))$ are not independent of c_i. For a related discussion, in the context of instrumental variables, see Athey and Stern (1998). Heckman, LaLonde, and Smith (2000) discuss alternative models that justify unconfoundedness. In these models individuals do attempt to optimize the same outcome that is the variable of interest to the evaluator. They show that selection-on-observables assumptions can be justified by imposing restrictions on the way individuals form their expectations about the unknown potential outcomes. In general, therefore, a researcher may wish to consider, either as a final analysis or as part of a larger investigation, estimates based on the unconfoundedness assumption.

Given the two key assumptions, unconfoundedness and overlap, one can identify the average treatment effects. The key insight is that given unconfoundedness, the following equalities hold:

$$\mu_w(x) = \mathbb{E}[Y(w)|X = x] = \mathbb{E}[Y(w)|W = w, X = x]$$

$$= \mathbb{E}[Y|W = w, X = x],$$

and thus $\mu_w(x)$ is identified. Thus one can estimate the average treatment effect τ by first estimating the average treatment effect for a subpopulation with covariates $X = x$:

$$\tau(x) \equiv \mathbb{E}[Y(1) - Y(0)|X = x] = \mathbb{E}[Y(1)|X = x]$$

$$- \mathbb{E}[Y(0)|X = x] = \mathbb{E}[Y(1)|X = x, W = 1]$$

$$- \mathbb{E}[Y(0)|X = x, W = 0] = \mathbb{E}[Y|X, W = 1]$$

$$- \mathbb{E}[Y|X, W = 0];$$

followed by averaging over the appropriate distribution of x. To make this feasible, one needs to be able to estimate the expectations $\mathbb{E}[Y|X = x, W = w]$ for all values of w and x in the support of these variables. This is where the second assumption enters. If the overlap assumption is violated at $X = x$, it would be infeasible to estimate both $\mathbb{E}[Y|X = x, W = 1]$ and $\mathbb{E}[Y|X = x, W = 0]$, because at those values of x there would be either only treated or only control units.

Some researchers use weaker versions of the unconfoundedness assumption (for example, Heckman, Ichimura, and Todd, 1998). If the interest is in the PATE, it is sufficient to assume that

ASSUMPTION 2.3 (MEAN INDEPENDENCE):

$$\mathbb{E}[Y(w)|W, X] = \mathbb{E}[Y(w)|X],$$

for $w = 0, 1$.

Although this assumption is unquestionably weaker, in practice it is rare that a convincing case is made for the weaker assumption 2.3 without the case being equally strong for the stronger version 2.1. The reason is that the weaker assumption is intrinsically tied to functional-form assumptions, and as a result one cannot identify average effects on transformations of the original outcome (such as logarithms) without the stronger assumption.

One can weaken the unconfoundedness assumption in a different direction if one is only interested in the average effect for the treated (see, for example, Heckman, Ichimura, & Todd, 1997). In that case one need only assume

ASSUMPTION 2.4 (UNCONFOUNDEDNESS FOR CONTROLS):

$$Y(0) \perp W|X.$$

and the weaker overlap assumption

ASSUMPTION 2.5 (WEAK OVERLAP):

$$\Pr(W = 1|X) < 1.$$

These two assumptions are sufficient for identification of PATT and SATT, because the moments of the distribution of $Y(1)$ for the treated are directly estimable.

An important result building on the unconfoundedness assumption shows that one need not condition simulta-

[3] If we are interested in the average effect for firms that did adopt the new technology (PATT), the following assumptions can be weakened slightly.

neously on all covariates. The following result shows that all biases due to observable covariates can be removed by conditioning solely on the propensity score:

Lemma 2.1 (Unconfoundedness Given the Propensity Score: Rosenbaum and Rubin, 1983a): Suppose that assumption 2.1 holds. Then

$$(Y(0),\ Y(1)) \perp W|e(X).$$

Proof: We will show that $\Pr(W = 1|Y(0),\ Y(1),\ e(X)) = \Pr(W = 1|e(X)) = e(X)$, implying independence of $(Y(0),\ Y(1))$ and W conditional on $e(X)$. First, note that

$$\begin{aligned}\Pr(W = 1|Y(0),\ Y(1),\ e(X)) &= \mathbb{E}[W = 1|Y(0),\ Y(1),\ e(X)]\\ &= \mathbb{E}[\mathbb{E}[W|Y(0),\ Y(1),\ e(X),\ X]|Y(0),\ Y(1),\ e(X)]\\ &= \mathbb{E}[\mathbb{E}[W|Y(0),\ Y(1),\ X]|Y(0),\ Y(1),\ e(X)]\\ &= \mathbb{E}[\mathbb{E}[W|X]|Y(0),\ Y(1),\ e(X)]\\ &= \mathbb{E}[e(X)|Y(0),\ Y(1),\ e(X)] = e(X),\end{aligned}$$

where the last equality follows from unconfoundedness. The same argument shows that

$$\begin{aligned}\Pr(W = 1|e(X)) &= \mathbb{E}[W = 1|e(X)] = \mathbb{E}[\mathbb{E}[W = 1|X]|e(X)]\\ &= \mathbb{E}[e(X)|e(X)] = e(X).\end{aligned}$$

■

Extensions of this result to the multivalued treatment case are given in Imbens (2000) and Lechner (2001). To provide intuition for Rosenbaum and Rubin's result, recall the textbook formula for omitted variable bias in the linear regression model. Suppose we have a regression model with two regressors:

$$Y_i = \beta_0 + \beta_1 \cdot W_i + \beta_2' X_i + \varepsilon_i.$$

The bias of omitting X from the regression on the coefficient on W is equal to $\beta_2'\delta$, where δ is the vector of coefficients on W in regressions of the elements of X on W. By conditioning on the propensity score we remove the correlation between X and W, because $X \perp W|e(X)$. Hence omitting X no longer leads to any bias (although it may still lead to some efficiency loss).

D. *Distributional and Quantile Treatment Effects*

Most of the literature has focused on estimating ATEs. There are, however, many cases where one may wish to estimate other features of the joint distribution of outcomes. Lehman (1974) and Doksum (1974) introduce quantile treatment effects as the difference in quantiles between the two marginal treated and control outcome distributions.[4] Bitler, Gelbach, and Hoynes (2002) estimate these in a randomized evaluation of a social program. In instrumental variables settings Abadie, Angrist, and Imbens (2002) and Chernozhukov and Hansen (2001) investigate estimation of differences in quantiles of the two marginal potential outcome distributions, either for the entire population or for subpopulations.

Assumptions 2.1 and 2.2 also allow for identification of the full marginal distributions of $Y(0)$ and $Y(1)$. To see this, first note that we can identify not just the average treatment effect $\tau(x)$, but also the averages of the two potential outcomes, $\mu_0(x)$ and $\mu_0(x)$. Second, by these assumptions we can similarly identify the averages of any function of the basic outcomes, $\mathbb{E}[\,g(Y(0))]$ and $\mathbb{E}[\,g(Y(1))]$. Hence we can identify the average values of the indicators $1\{Y(0) \le y\}$ and $1\{Y(1) \le y\}$, and thus the distribution function of the potential outcomes at y. Given identification of the two distribution functions, it is clear that one can also identify quantiles of the two potential outcome distributions. Firpo (2002) develops an estimator for such quantiles under unconfoundedness.

E. *Efficiency Bounds and Asymptotic Variances for Population-Average Treatment Effects*

Next I review some results on the efficiency bound for estimators of the ATEs τ^P, and τ_T^P. This requires both the assumptions of unconfoundedness and overlap (Assumptions 2.1 and 2.2) and some smoothness assumptions on the conditional expectations of potential outcomes and the treatment indicator (for details, see Hahn, 1998). Formally, Hahn (1998) shows that for any regular estimator for τ^P, denoted by $\hat{\tau}$, with

$$\sqrt{N} \cdot (\hat{\tau} - \tau^P) \xrightarrow{d} \mathcal{N}(0, V),$$

it must be that

$$V \ge \mathbb{E}\left[\frac{\sigma_1^2(X)}{e(X)} + \frac{\sigma_0^2(X)}{1 - e(X)} + (\tau(X) - \tau^P)^2\right].$$

Knowledge of the propensity score does not affect this efficiency bound.

Hahn also shows that asymptotically linear estimators exist with such variance, and hence such efficient estimators can be approximated as

$$\hat{\tau} = \tau^P + \frac{1}{N}\sum_{i=1}^{N} \psi(Y_i, W_i, X_i, \tau^P) + o_p(N^{-1/2}),$$

where $\psi(\cdot)$ is the efficient score:

[4] In contrast, Heckman, Smith, and Clemens (1997) focus on estimation of bounds on the joint distribution of $(Y(0),\ Y(1))$. One cannot without strong untestable assumptions identify the full joint distribution, since one can never observe both potential outcomes simultaneously, but one can nevertheless derive bounds on functions of the two distributions.

$$\psi(y, w, x, \tau^P) = \left(\frac{wy}{e(x)} - \frac{(1-w)y}{1-e(x)}\right) - \tau^P - \left(\frac{\mu_1(x)}{e(x)} + \frac{\mu_0(x)}{1-e(x)}\right)[w - e(x)]. \quad (1)$$

Hahn (1998) also reports the efficiency bound for τ_T^P, both with and without knowledge of the propensity score. For τ_T^P the efficiency bound given knowledge of $e(X)$ is

$$\mathbb{E}\left[\frac{e(X)\ \mathrm{Var}(Y(1)|X)}{\mathbb{E}[e(X)]^2} + \frac{e(X)^2\ \mathrm{Var}(Y(0)|X)}{\mathbb{E}[e(X)]^2(1-e(X))} + (\tau(X) - \tau_T^P)^2 \frac{e(X)^2}{\mathbb{E}[e(X)]^2}\right].$$

If the propensity score is not known, unlike the bound for τ^P, the efficiency bound for τ_T^P is affected. For τ_T^P the bound without knowledge of the propensity score is

$$\mathbb{E}\left[\frac{e(X)\ \mathrm{Var}(Y(1)|X)}{\mathbb{E}[e(X)]^2} + \frac{e(X)^2\ \mathrm{Var}(Y(0)|X)}{\mathbb{E}[e(X)]^2(1-e(X))} + (\tau(X) - \tau_T^P)^2 \frac{e(X)}{\mathbb{E}[e(X)]^2}\right],$$

which is higher by

$$\mathbb{E}\left[(\tau(X) - \tau_T^P)^2 \cdot \frac{e(X)(1-e(X))}{\mathbb{E}[e(X)]^2}\right].$$

The intuition that knowledge of the propensity score affects the efficiency bound for the average effect for the treated (PATT), but not for the overall average effect (PATE), goes as follows. Both are weighted averages of the treatment effect conditional on the covariates, $\tau(x)$. For the PATE the weight is proportional to the density of the covariates, whereas for the PATT the weight is proportional to the product of the density of the covariates and the propensity score (see, for example, Hirano, Imbens, and Ridder, 2003). Knowledge of the propensity score implies one does not need to estimate the weight function and thus improves precision.

F. *Efficiency Bounds and Asymptotic Variances for Conditional and Sample Average Treatment Effects*

Consider the leading term of the efficient estimator for the PATE, $\tilde{\tau} = \tau^P + \bar{\psi}$, where $\bar{\psi} = (1/N)\sum \psi(Y_i, W_i, X_i, \tau^P)$, and let us view this as an estimator for the SATE, instead of as an estimator for the PATE. I will show that, first, this estimator is unbiased, conditional on the covariates and the potential outcomes, and second, it has lower variance as an estimator of the SATE than as an estimator of the PATE. To see that the estimator is unbiased note that with the efficient score $\psi(y, w, x, \tau)$ given in equation (1),

$$\mathbb{E}[\psi(Y, W, X, \tau^P)|Y(0), Y(1), X)] = Y(1) - Y(0) - \tau^P,$$

and thus

$$\mathbb{E}[\tilde{\tau}|(Y_i(0), Y_i(1), X_i)_{i=1}^N] = \mathbb{E}[\bar{\psi}] + \tau^P = \frac{1}{N}\sum_{i=1}^N (Y_i(1) - Y_i(0)).$$

Hence

$$\mathbb{E}[\tilde{\tau} - \tau^S|(Y_i(0), Y_i(1), X_i)_{i=1}^N] = \frac{1}{N}\sum_{i=1}^N (Y_i(1) - Y_i(0)) - \tau^S = 0.$$

Next, consider the normalized variance:

$$V^P = N \cdot \mathbb{E}[(\tilde{\tau} - \tau^S)^2] = N \cdot \mathbb{E}[(\bar{\psi} + \tau^P - \tau^S)^2].$$

Note that the variance of $\tilde{\tau}$ as an estimator of τ^P can be expressed, using the fact that $\psi(\cdot)$ is the efficient score, as

$$N \cdot \mathbb{E}[(\tilde{\tau} - \tau^P)^2] = N \cdot \mathbb{E}[(\bar{\psi})^2] = N \cdot \mathbb{E}[(\bar{\psi}(Y, W, X, \tau^P) + (\tau^P - \tau^S) - (\tau^P - \tau^S))^2].$$

Because

$$\mathbb{E}[(\bar{\psi}(Y, W, X, \tau^P) + (\tau^P - \tau^S)) \cdot (\tau^P - \tau^S)] = 0$$

[as follows by using iterated expectations, first conditioning on X, $Y(0)$, and $Y(1)$], it follows that

$$N \cdot \mathbb{E}[(\tilde{\tau} - \tau^P)^2] = N \cdot \mathbb{E}[(\tilde{\tau} - \tau^S)^2] + N \cdot \mathbb{E}[(\tau^S - \tau^P)^2] = N \cdot \mathbb{E}[(\tilde{\tau} - \tau^S)^2] + N \cdot \mathbb{E}[(Y(1) - Y(0) - \tau^P)^2].$$

Thus, the same statistic that as an estimator of the population average treatment effect τ^P has a normalized variance equal to V^P, as an estimator of τ^S has the property

$$\sqrt{N}(\tilde{\tau} - \tau^S) \xrightarrow{d} \mathcal{N}(0, V^S),$$

with

$$V^S = V^P - \mathbb{E}[(Y(1) - Y(0) - \tau^P)^2].$$

As an estimator of τ^S the variance of $\tilde{\tau}$ is lower than its variance as an estimator of τ^P, with the difference equal to the variance of the treatment effect.

The same line of reasoning can be used to show that

$$\sqrt{N}(\tilde{\tau} - \overline{\tau(X)}) \xrightarrow{d} \mathcal{N}(0, V^{\overline{\tau(X)}}),$$

with

$$V^{\overline{\tau(X)}} = V^P - \mathbb{E}[(\tau(X) - \tau^P)^2],$$

and

$$V^S = V^{\overline{\tau(X)}} - \mathbb{E}[(Y(1) - Y(0) - \tau(X))^2].$$

An example to illustrate these points may be helpful. Suppose that $X \in \{0, 1\}$, with $\Pr(X = 1) = p_x$ and $\Pr(W = 1|X) = 1/2$. Suppose that $\tau(x) = 2x - 1$, and $\sigma^2_w(x)$ is very small for all x and w. In that case the average treatment effect is $p_x \cdot 1 + (1 - p_x) \cdot (-1) = 2p_x - 1$. The efficient estimator in this case, assuming only unconfoundedness, requires separately estimating $\tau(x)$ for $x = 0$ and 1, and averaging these two by the empirical distribution of X. The variance of $\sqrt{N}(\hat{\tau} - \tau^S)$ will be small because $\sigma^2_w(x)$ is small, and according to the expressions above, the variance of $\sqrt{N}(\tau - \tau^P)$ will be larger by $4p_x(1 - p_x)$. If p_x differs from 1/2, and so PATE differs from 0, the confidence interval for PATE in small samples will tend to include zero. In contrast, with $\sigma^2_w(x)$ small enough and N odd [and both N_0 and N_1 at least equal to 2, so that one can estimate $\sigma^2_w(x)$], the standard confidence interval for τ^S will exclude 0 with probability 1. The intuition is that τ^P is much more uncertain because it depends on the distribution of the covariates, whereas the uncertainty about τ^S depends only on the conditional outcome variances and the propensity score.

The difference in asymptotic variances raises the issue of how to estimate the variance of the sample average treatment effect. Specific estimators for the variance will be discussed in section IV, but here I will introduce some general issues surrounding their estimation. Because the two potential outcomes for the same unit are never observed simultaneously, one cannot directly infer the variance of the treatment effect. This is the same issue as the nonidentification of the correlation coefficient. One can, however, estimate a lower bound on the variance of the treatment effect, leading to an upper bound on the variance of the estimator of the SATE, which is equal to $V^{\overline{\tau(X)}}$. Decomposing the variance as

$$\begin{aligned}\mathbb{E}[(Y(1) - Y(0) - \tau^P)^2] &= \mathbb{V}(\mathbb{E}[Y(1) - Y(0) - \tau^P|X]) \\ &\quad + \mathbb{E}[\mathbb{V}(Y(1) - Y(0) - \tau^P|X)], \\ &= \mathbb{V}(\tau(X) - \tau^P) + \mathbb{E}[\sigma^2_1(X) + \sigma^2_0(X) \\ &\quad - 2\rho(X)\sigma_0(X)\sigma_1(X)],\end{aligned}$$

we can consistently estimate the first term, but generally say little about the second other than that it is nonnegative. One can therefore bound the variance of $\tilde{\tau} - \tau^S$ from above by

$$\begin{aligned}&\mathbb{E}[\psi(Y, W, X, \tau^P)^2] - \mathbb{E}[(Y(1) - Y(0)) - \tau^P)^2] \\ &\quad \leq \mathbb{E}[\psi(Y, W, X, \tau^P)^2] - \mathbb{E}[(\tau(X) - \tau^P)^2] \\ &\quad = \mathbb{E}\left[\frac{\sigma^2_1(X)}{e(X)} + \frac{\sigma^2_0(X)}{1 - e(X)}\right] = V^{\overline{\tau(X)}},\end{aligned}$$

and use this upper-bound variance estimate to construct confidence intervals that are guaranteed to be conservative. Note the connection with Neyman's (1923) discussion of conservative confidence intervals for average treatment effects in experimental settings. It should be noted that the difference between these variances is of the same order as the variance itself, and therefore not a small-sample problem. Only when the treatment effect is known to be constant can it be ignored. Depending on the correlation between the outcomes and the covariates, this may change the standard errors considerably. It should also be noted that bootstrapping methods in general lead to estimation of $\mathbb{E}[(\tilde{\tau} - \tau^P)^2]$, rather than $\mathbb{E}[(\tilde{\tau} - \tau(X))^2]$, which are generally too big.

III. Estimating Average Treatment Effects

There have been a number of statistics proposed for estimating the PATE and PATT, all of which are also appropriate estimators of the sample versions (SATE and SATT) and the conditional average versions (CATE and CATT). (The implications of focusing on SATE or CATE rather than PATE only arise when estimating the variance, and so I will return to this distinction in section IV. In the current section all discussion applies equally to all estimands.) Here I review some of these estimators, organized into five groups.

The first set, referred to as *regression* estimators, consists of methods that rely on consistent estimation of the two conditional regression functions, $\mu_0(x)$ and $\mu_1(x)$. These estimators differ in the way that they estimate these elements, but all rely on estimators that are consistent for these regression functions.

The second set, *matching* estimators, compare outcomes across pairs of matched treated and control units, with each unit matched to a fixed number of observations with the opposite treatment. The bias of these within-pair estimates of the average treatment effect disappears as the sample size increases, although their variance does not go to zero, because the number of matches remains fixed.

The third set of estimators is characterized by a central role for the propensity score. Four leading approaches in this set are weighting by the reciprocal of the propensity score, blocking on the propensity score, regression on the propensity score, and matching on the propensity score.

The fourth category consists of estimators that rely on a combination of these methods, typically combining regression with one of its alternatives. The motivation for these combinations is that although in principle any one of these methods can remove all of the bias associated with the covariates, combining two may lead to more robust inference. For example, matching leads to consistent estimators for average treatment effects under weak conditions, so matching and regression can combine some of the desirable variance properties of regression with the consistency of matching. Similarly, a combination of weighting and regression, using parametric models for both the propensity score

and the regression functions, can lead to an estimator that is consistent even if only one of the models is correctly specified ("doubly robust" in the terminology of Robins & Ritov, 1997).

Finally, in the fifth group I will discuss Bayesian approaches to inference for average treatment effects.

Only some of the estimators discussed below achieve the semiparametric efficiency bound, yet this does not mean that these should necessarily be preferred in practice—that is, in finite samples. More generally, the debate concerning the practical advantages of the various estimators, and the settings in which some are more attractive than others, is still ongoing, with as of yet no firm conclusions. Although all estimators, either implicitly or explicitly, estimate the two unknown regression functions or the propensity score, they do so in very different ways. Differences in smoothness of the regression function or the propensity score, or relative discreteness of the covariates in specific applications, may affect the relative attractiveness of the estimators.

In addition, even the appropriateness of the standard asymptotic distributions as a guide towards finite-sample performance is still debated (see, for example, Robins & Ritov, 1997, and Angrist & Hahn, 2004). A key feature that casts doubt on the relevance of the asymptotic distributions is that the $\sqrt{N}$ consistency is obtained by averaging a nonparametric estimator of a regression function that itself has a slow nonparametric convergence rate over the empirical distribution of its argument. The dimension of this argument affects the rate of convergence for the unknown function [the regression functions $\mu_w(x)$ or the propensity score $e(x)$], but not the rate of convergence for the estimator of the parameter of interest, the average treatment effect. In practice, however, the resulting approximations of the ATE can be poor if the argument is of high dimension, in which case information about the propensity score is of particular relevance. Although Hahn (1998) showed, as discussed above, that for the standard asymptotic distributions knowledge of the propensity score is irrelevant (and conditioning only on the propensity score is in fact less efficient than conditioning on all covariates), conditioning on the propensity score involves only one-dimensional nonparametric regression, suggesting that the asymptotic approximations may be more accurate. In practice, knowledge of the propensity score may therefore be very informative.

Another issue that is important in judging the various estimators is how well they perform when there is only limited overlap in the covariate distributions of the two treatment groups. If there are regions in the covariate space with little overlap (propensity score close to 0 or 1), ATE estimators should have relatively high variance. However, this is not always the case for estimators based on tightly parametrized models for the regression functions, where outliers in covariate values can lead to spurious precision for regression parameters. Regions of small overlap can also be difficult to detect directly in high-dimensional covariate spaces, as they can be masked for any single variable.

A. *Regression*

The first class of estimators relies on consistent estimation of $\mu_w(x)$ for $w = 0, 1$. Given $\hat{\mu}_w(x)$ for these regression functions, the PATE, SATE, and CATE are estimated by averaging their differences over the empirical distribution of the covariates:

$$\hat{\tau}_{\text{reg}} = \frac{1}{N}\sum_{i=1}^{N}[\hat{\mu}_1(X_i) - \hat{\mu}_0(X_i)]. \tag{2}$$

In most implementations the average of the predicted treated outcome for the treated is equal to the average observed outcome for the treated [so that $\sum_i W_i \cdot \hat{\mu}_1(X_i) = \sum_i W_i \cdot Y_i$], and similarly for the controls, implying that $\hat{\tau}_{\text{reg}}$ can also be written as

$$\frac{1}{N}\sum_{i=1}^{N} W_i \cdot [Y_i - \hat{\mu}_0(X_i)] + (1 - W_i) \cdot [\hat{\mu}_1(X_i) - Y_i].$$

For the PATT and SATT typically only the control regression function is estimated; we only need predict the outcome under the control treatment for the treated units. The estimator then averages the difference between the actual outcomes for the treated and their estimated outcomes under the control:

$$\hat{\tau}_{\text{reg},T} = \frac{1}{N_T}\sum_{i=1}^{N} W_i \cdot [Y_i - \hat{\mu}_0(X_i)]. \tag{3}$$

Early estimators for $\mu_w(x)$ included parametric regression functions—for example, linear regression (as in Rubin, 1977). Such parametric alternatives include least squares estimators with the regression function specified as

$$\mu_w(x) = \beta' x + \tau \cdot w,$$

in which case the average treatment effect is equal to τ. In this case one can estimate τ directly by least squares estimation using the regression function

$$Y_i = \alpha + \beta' X_i + \tau \cdot W_i + \varepsilon_i.$$

More generally, one can specify separate regression functions for the two regimes:

$$\mu_w(x) = \beta_w' x.$$

In that case one can estimate the two regression functions separately on the two subsamples and then substitute the predicted values in equation (2). These simple regression estimators may be very sensitive to differences in the covariate distributions for treated and control units. The

reason is that in that case the regression estimators rely heavily on extrapolation. To see this, note that the regression function for the controls, $\mu_0(x)$, is used to predict missing outcomes for the treated. Hence on average one wishes to predict the control outcome at $\bar{X}_T$, the average covariate value for the treated. With a linear regression function, the average prediction can be written as $\bar{Y}_C + \hat{\beta}'(\bar{X}_T - \bar{X}_C)$. With $\bar{X}_T$ very close to the average covariate value for the controls, $\bar{X}_C$, the precise specification of the regression function will not matter very much for the average prediction. However, with the two averages very different, the prediction based on a linear regression function can be very sensitive to changes in the specification.

More recently, nonparametric estimators have been proposed. Hahn (1998) recommends estimating first the three conditional expectations $g_1(x) = \mathbb{E}[WY|X]$, $g_0(x) = \mathbb{E}[(1 - W)Y|X]$, and $e(x) = \mathbb{E}[W|X]$ nonparametrically using series methods. He then estimates $\mu_w(x)$ as

$$\hat{\mu}_1(x) = \frac{\hat{g}_1(x)}{\hat{e}(x)}, \quad \hat{\mu}_0(x) = \frac{\hat{g}_0(x)}{1 - \hat{e}(x)},$$

and shows that the estimators for both PATE and PATT achieve the semiparametric efficiency bounds discussed in section IIE (the latter even when the propensity score is unknown).

Using this series approach, however, it is unnecessary to estimate all three of these conditional expectations ($\mathbb{E}[YW|X]$, $\mathbb{E}[Y(1 - W)|X]$, and $\mathbb{E}[W|X]$) to estimate $\mu_w(x)$. Instead one can use series methods to directly estimate the two regression functions $\mu_w(x)$, eliminating the need to estimate the propensity score (Imbens, Newey, and Ridder, 2003).

Heckman, Ichimura, and Todd (1997, 1998) and Heckman, Ichimura, Smith, and Todd (1998) consider kernel methods for estimating $\mu_w(x)$, in particular focusing on local linear approaches. The simple kernel estimator has the form

$$\hat{\mu}_w(x) = \sum_{i:W_i=w} Y_i \cdot K\left(\frac{X_i - x}{h}\right) \Bigg/ \sum_{i:W_i=w} K\left(\frac{X_i - x}{h}\right),$$

with a kernel $K(\cdot)$ and bandwidth h. In the local linear kernel regression the regression function $\mu_w(x)$ is estimated as the intercept β_0 in the minimization problem

$$\min_{\beta_0,\beta_1} \sum_{i:W_i=w} [Y_i - \beta_0 - \beta_1'(X_i - x)]^2 \cdot K\left(\frac{X_i - x}{h}\right).$$

In order to control the bias of their estimators, Heckman, Ichimura, and Todd (1998) require that the order of the kernel be at least as large as the dimension of the covariates. That is, they require the use of a kernel function $K(z)$ such that $\int_z z^r K(z)\, dz = 0$ for $r \leq \dim(X)$, so that the kernel must be negative on part of the range, and the implicit averaging involves negative weights. We shall see this role of the dimension of the covariates again for other estimators.

For the average treatment effect for the treated (PATT), it is important to note that with the propensity score known, the estimator given in equation (3) is generally not efficient, irrespective of the estimator for $\mu_0(x)$. Intuitively, this is because with the propensity score known, the average $\sum W_i Y_i / N_T$ is not efficient for the population expectation $\mathbb{E}[Y(1)|W = 1]$. An efficient estimator (as in Hahn, 1998) can be obtained by weighting all the estimated treatment effects, $\hat{\mu}_1(X_i) - \hat{\mu}_0(X_i)$, by the probability of receiving the treatment:

$$\hat{\tau}_{\text{reg},T} = \sum_{i=1}^{N} e(X_i) \cdot [\hat{\mu}_1(X_i) - \hat{\mu}_0(X_i)] \Bigg/ \sum_{i=1}^{N} e(X_i). \tag{4}$$

In other words, instead of estimating $\mathbb{E}[Y(1)|W = 1]$ as $\sum W_i Y_i / N_T$ using only the treated observations, it is estimated using all units, as $\sum \hat{\mu}_1(X_i) \cdot e(X_i) / \sum e(X_i)$. Knowledge of the propensity score improves the accuracy because it allows one to exploit the control observations to adjust for imbalances in the sampling of the covariates.

For all of the estimators in this section an important issue is the choice of the smoothing parameter. In Hahn's case, after choosing the form of the series and the sequence, the smoothing parameter is the number of terms in the series. In Heckman, Ichimura, and Todd's case it is the bandwidth of the kernel chosen. The evaluation literature has been largely silent concerning the optimal choice of the smoothing parameters, although the larger literature on nonparametric estimation of regression functions does provide some guidance, offering data-driven methods such as cross-validation criteria. The optimality properties of these criteria, however, are for estimation of the entire function, in this case $\mu_w(x)$. Typically the focus is on mean-integrated-squared-error criteria of the form $\int_x [\hat{\mu}_w(x) - \mu_w(x)]^2 f_X(x)\, dx$, with possibly an additional weight function. In the current problem, however, one is interested specifically in the average treatment effect, and so such criteria are not necessarily optimal. In particular, global smoothing parameters may be inappropriate, because they can be driven by the shape of the regression function and distribution of covariates in regions that are not important for the average treatment effect of interest. LaLonde's (1986) data set is a well-known example of this where much of probability mass of the nonexperimental control group is in a region with moderate to high earnings where few of the treated group are located. There is little evidence whether results for average treatment effects are more or less sensitive to the choice of smoothing parameter than results for estimation of the regression functions themselves.

B. *Matching*

As seen above, regression estimators impute the missing potential outcomes using the estimated regression function.

Thus, if $W_i = 1$, then $Y_i(1)$ is observed and $Y_i(0)$ is missing and imputed with a consistent estimator $\hat{\mu}_0(X_i)$ for the conditional expectation. Matching estimators also impute the missing potential outcomes, but do so using only the outcomes of nearest neighbors of the opposite treatment group. In that respect matching is similar to nonparametric kernel regression methods, with the number of neighbors playing the role of the bandwidth in the kernel regression. A formal difference is that the asymptotic distribution is derived conditional on the implicit bandwidth, that is, the number of neighbors, which is often fixed at one. Using such asymptotics, the implicit estimate $\hat{\mu}_w(x)$ is (close to) unbiased, but not consistent for $\mu_w(x)$. In contrast, the regression estimators discussed in the previous section rely on the consistency of $\mu_w(x)$.

Matching estimators have the attractive feature that given the matching metric, the researcher only has to choose the number of matches. In contrast, for the regression estimators discussed above, the researcher must choose smoothing parameters that are more difficult to interpret: either the number of terms in a series or the bandwidth in kernel regression. Within the class of matching estimators, using only a single match leads to the most credible inference with the least bias, at most sacrificing some precision. This can make the matching estimator easier to use than those estimators that require more complex choices of smoothing parameters, and may explain some of its popularity.

Matching estimators have been widely studied in practice and theory (for example, Gu & Rosenbaum, 1993; Rosenbaum, 1989, 1995, 2002; Rubin, 1973b, 1979; Heckman, Ichimura, & Todd, 1998; Dehejia & Wahba, 1999; Abadie & Imbens, 2002). Most often they have been applied in settings with the following two characteristics: (i) the interest is in the average treatment effect for the treated, and (ii) there is a large reservoir of potential controls. This allows the researcher to match each treated unit to one or more distinct controls (referred to as matching without replacement). Given the matched pairs, the treatment effect within a pair is then estimated as the difference in outcomes, with an estimator for the PATT obtained by averaging these within-pair differences. Since the estimator is essentially the difference between two sample means, the variance is calculated using standard methods for differences in means or methods for paired randomized experiments. The remaining bias is typically ignored in these studies. The literature has studied fast algorithms for matching the units, as fully efficient matching methods are computationally cumbersome (see, for example, Gu and Rosenbaum, 1993; Rosenbaum, 1995). Note that in such matching schemes the order in which the units are matched may be important.

Abadie and Imbens (2002) study both bias and variance in a more general setting where both treated and control units are (potentially) matched and matching is done with replacement (as in Dehejia & Wahba, 1999). The Abadie-Imbens estimator is implemented in Matlab and Stata (see Abadie et al., 2003).[5] Formally, given a sample, $\{(Y_i, X_i, W_i)\}_{i=1}^N$, let $\ell_m(i)$ be the index l that satisfies $W_l \neq W_i$ and

$$\sum_{j|W_j \neq W_i} 1\{\|X_j - X_i\| \leq \|X_l - X_i\|\} = m,$$

where $1\{\cdot\}$ is the indicator function, equal to 1 if the expression in brackets is true and 0 otherwise. In other words, $\ell_m(i)$ is the index of the unit in the opposite treatment group that is the m^{th} closest to unit i in terms of the distance measure based on the norm $\|\cdot\|$. In particular, $\ell_1(i)$ is the nearest match for unit i. Let $\mathcal{J}_M(i)$ denote the set of indices for the first M matches for unit i: $\mathcal{J}_M(i) = \{\ell_1(i), \ldots, \ell_M(i)\}$. Define the imputed potential outcomes as

$$\hat{Y}_i(0) = \begin{cases} Y_i & \text{if } W_i = 0, \\ \dfrac{1}{M} \displaystyle\sum_{j \in \mathcal{J}_M(i)} Y_j & \text{if } W_i = 1, \end{cases}$$

and

$$\hat{Y}_i(1) = \begin{cases} \dfrac{1}{M} \displaystyle\sum_{j \in \mathcal{J}_M(i)} Y_j & \text{if } W_i = 0, \\ Y_i & \text{if } W_i = 1. \end{cases}$$

The simple matching estimator discussed by Abadie and Imbens is then

$$\hat{\tau}_M^{\text{sm}} = \frac{1}{N} \sum_{i=1}^{N} [\hat{Y}_i(1) - \hat{Y}_i(0)]. \tag{5}$$

They show that the bias of this estimator is $O(N^{-1/k})$, where k is the dimension of the covariates. Hence, if one studies the asymptotic distribution of the estimator by normalizing by $\sqrt{N}$ [as can be justified by the fact that the variance of the estimator is $O(1/N)$], the bias does not disappear if the dimension of the covariates is equal to 2, and will dominate the large sample variance if k is at least 3.

Let me make clear three caveats to Abadie and Imbens's result. First, it is only the continuous covariates that should be counted in this dimension, k. With discrete covariates the matching will be exact in large samples; therefore such covariates do not contribute to the order of the bias. Second, if one matches only the treated, and the number of potential controls is much larger than the number of treated units, one can justify ignoring the bias by appealing to an asymptotic sequence where the number of potential controls increases faster than the number of treated units. Specifically, if the number of controls, N_0, and the number of treated, N_1, satisfy $N_1/N_0^{4/k} \to 0$, then the bias disappears in large samples after normalization by $\sqrt{N_1}$. Third, even though

[5] See Becker and Ichino (2002) and Sianesi (2001) for alternative Stata implementations of estimators for average treatment effects.

the order of the bias may be high, the actual bias may still be small if the coefficients in the leading term are small. This is possible if the biases for different units are at least partially offsetting. For example, the leading term in the bias relies on the regression function being nonlinear, and the density of the covariates having a nonzero slope. If one of these two conditions is at least close to being satisfied, the resulting bias may be fairly limited. To remove the bias, Abadie and Imbens suggest combining the matching process with a regression adjustment, as I will discuss in section IIID.

Another point made by Abadie and Imbens is that matching estimators are generally not efficient. Even in the case where the bias is of low enough order to be dominated by the variance, the estimators are not efficient given a fixed number of matches. To reach efficiency one would need to increase the number of matches with the sample size. If $M \to \infty$, with $M/N \to 0$, then the matching estimator is essentially like a regression estimator, with the imputed missing potential outcomes consistent for their conditional expectations. However, the efficiency gain of such estimators is of course somewhat artificial. If in a given data set one uses M matches, one can calculate the variance as if this number of matches increased at the appropriate rate with the sample size, in which case the estimator would be efficient, or one could calculate the variance conditional on the number of matches, in which case the same estimator would be inefficient. Little is yet known about the optimal number of matches, or about data-dependent ways of choosing this number.

In the above discussion the distance metric in choosing the optimal matches was the standard Euclidean metric:

$$d_E(x, z) = (x - z)'(x - z).$$

All of the distance metrics used in practice standardize the covariates in some manner. Abadie and Imbens use the diagonal matrix of the inverse of the covariate variances:

$$d_{AI}(x, z) = (x - z)' \text{ diag } (\Sigma_X^{-1}) (x - z),$$

where Σ_X is the covariance matrix of the covariates. The most common choice is the Mahalanobis metric (see, for example, Rosenbaum and Rubin, 1985), which uses the inverse of the covariance matrix of the pretreatment variables:

$$d_M(x, z) = (x - z)'\Sigma_X^{-1}(x - z).$$

This metric has the attractive property that it reduces differences in covariates within matched pairs in all directions.[6] See for more formal discussions Rubin and Thomas (1992).

Zhao (2004), in an interesting discussion of the choice of metrics, suggests some alternatives that depend on the correlation between covariates, treatment assignment, and outcomes. He starts by assuming that the propensity score has a logistic form

$$e(x) = \frac{\exp(x'\gamma)}{1 + \exp(x'\gamma)},$$

and that the regression functions are linear:

$$\mu_w(x) = \alpha_w + x'\beta.$$

He then considers two alternative metrics. The first weights absolute differences in the covariates by the coefficient in the propensity score:

$$d_{Z1}(x, z) = \sum_{k=1}^{K} |x_k - z_k| \cdot |\gamma_k|,$$

and the second weights them by the coefficients in the regression function:

$$d_{Z2}(x, z) = \sum_{k=1}^{K} |x_k - z_k| \cdot |\beta_k|,$$

where x_k and z_k are the k^{th} elements of the K-dimensional vectors x and z respectively.

In light of this discussion, it is interesting to consider optimality of the metric. Suppose, following Zhao (2004), that the regression functions are linear with coefficients β_w. Now consider a treated unit with covariate vector x who will be matched to a control unit with covariate vector z. The bias resulting from such a match is $(z - x)'\beta_0$. If one is interested in minimizing for each match the squared bias, one should choose the first match by minimizing over the control observations $(z - x)'\beta_0\beta_0'(z - x)$. Yet typically one does not know the value of the regression coefficients, in which case one may wish to minimize the expected squared bias. Using a normal distribution for the regression errors, and a flat prior on β_0, the posterior distribution for β_0 is normal with mean $\hat{\beta}_0$ and variance $\Sigma_X^{-1}\sigma^2/N$. Hence the expected squared bias from a match is

[6] However, using the Mahalanobis metric can also have less attractive implications. Consider the case where one matches on two highly correlated covariates, X_1 and X_2 with equal variances. For specificity, suppose that the correlation coefficient is 0.9 and both variances are 1. Suppose that we wish to match a treated unit i with $X_{i1} = X_{i2} = 0$. The two potential matches are unit j with $X_{j1} = X_{j2} = 5$ and unit k with $X_{k1} = 4$ and $X_{k2} = 0$. The difference in covariates for the first match is the vector $(5, 5)'$, and the difference in covariates for the second match is $(4, 0)'$. Intuitively it may seem that the second match is better: it is strictly closer to the treated unit than the first match for both covariates. Using the Abadie-Imbens metric diag (Σ_X^{-1}), this is in fact true. Under that metric the distance between the second match and the treated unit is 16, considerably smaller than 50, the distance between the first match and the treated unit. Using the Mahalanobis metric, however, the distance for the first match is 26, and the distance for the second match is much higher at 84. Because of the correlation between the covariates in the sample, the difference between the matches is interpreted very differently under the two metrics. To choose between the standard and the Mahalanobis metric one needs to consider what the appropriate match would be in this case.

$$\mathbb{E}[(z-x)'\beta_0\beta_0'(z-x)] = (z-x)'(\hat{\beta}_0\hat{\beta}_0' + \sigma^2\Sigma_X^{-1}/N) \times (z-x).$$

In this argument the optimal metric is a combination of the sample covariance matrix plus the outer product of the regression coefficients, with the former scaled down by a factor $1/N$:

$$d^*(z,x) = (z-x)'(\hat{\beta}_w\hat{\beta}_w' + \sigma_w^2\Sigma_{X,w}^{-1}/N)(z-x).$$

A clear problem with this approach is that when the regression function is misspecified, matching with this particular metric may not lead to a consistent estimator. On the other hand, when the regression function is correctly specified, it would be more efficient to use the regression estimators than any matching approach. In practice one may want to use a metric that combines some of the optimal weighting with some safeguards in case the regression function is misspecified.

So far there is little experience with any alternative metrics beyond the Mahalanobis metric. Zhao (2004) reports the results of some simulations using his proposed metrics, finding no clear winner given his specific design, although his findings suggest that using the outcomes in defining the metric is a promising approach.

C. *Propensity Score Methods*

Since the work by Rosenbaum and Rubin (1983a) there has been considerable interest in methods that avoid adjusting directly for all covariates, and instead focus on adjusting for differences in the propensity score, the conditional probability of receiving the treatment. This can be implemented in a number of different ways. One can weight the observations using the propensity score (and indirectly also in terms of the covariates) to create balance between treated and control units in the weighted sample. Hirano, Imbens, and Ridder (2003) show how such estimators can achieve the semiparametric efficiency bound. Alternatively one can divide the sample into subsamples with approximately the same value of the propensity score, a technique known as blocking. Finally, one can directly use the propensity score as a regressor in a regression approach.

In practice there are two important cases. First, suppose the researcher knows the propensity score. In that case all three of these methods are likely to be effective in eliminating bias. Even if the resulting estimator is not fully efficient, one can easily modify it by using a parametric estimate of the propensity score to capture most of the efficiency loss. Furthermore, since these estimators do not rely on high-dimensional nonparametric regression, this suggests that their finite-sample properties are likely to be relatively attractive.

If the propensity score is not known, the advantages of the estimators discussed below are less clear. Although they avoid the high-dimensional nonparametric regression of the two conditional expectations $\mu_w(x)$, they require instead the equally high-dimensional nonparametric regression of the treatment indicator on the covariates. In practice the relative merits of these estimators will depend on whether the propensity score is more or less smooth than the regression functions, and on whether additional information is available about either the propensity score or the regression functions.

Weighting: The first set of propensity-score estimators use the propensity scores as weights to create a balanced sample of treated and control observations. Simply taking the difference in average outcomes for treated and controls,

$$\hat{\tau} = \frac{\sum W_iY_i}{\sum W_i} - \frac{\sum (1-W_i)Y_i}{\sum 1-W_i},$$

is not unbiased for $\tau^P = \mathbb{E}[Y(1) - Y(0)]$, because, conditional on the treatment indicator, the distributions of the covariates differ. By weighting the units by the reciprocal of the probability of receiving the treatment, one can undo this imbalance. Formally, weighting estimators rely on the equalities

$$\mathbb{E}\left[\frac{WY}{e(X)}\right] = \mathbb{E}\left[\frac{WY(1)}{e(X)}\right] = \mathbb{E}\left[\mathbb{E}\left[\frac{WY(1)}{e(X)}\middle| X\right]\right] = \mathbb{E}\left[\frac{e(X)\cdot\mathbb{E}[Y(1)|X]}{e(X)}\right] = \mathbb{E}[Y(1)],$$

using unconfoundedness in the second to last equality, and similarly

$$\mathbb{E}\left[\frac{(1-W)Y}{1-e(X)}\right] = \mathbb{E}[Y(0)],$$

implying

$$\tau^P = \mathbb{E}\left[\frac{W\cdot Y}{e(X)} - \frac{(1-W)\cdot Y}{1-e(X)}\right].$$

With the propensity score known one can directly implement this estimator as

$$\tilde{\tau} = \frac{1}{N}\sum_{i=1}^{N}\left(\frac{W_iY_i}{e(X_i)} - \frac{(1-W_i)Y_i}{1-e(X_i)}\right). \tag{6}$$

In this particular form this is not necessarily an attractive estimator. The main reason is that, although the estimator can be written as the difference between a weighted average of the outcomes for the treated units and a weighted average of the outcomes for the controls, the weights do not necessarily add to 1. Specifically, in equation (6), the weights for the treated units add up to $[\sum W_i/e(X_i)]/N$. In expectation this is equal to 1, but because its variance is positive, in any

given sample some of the weights are likely to deviate from 1.

One approach for improving this estimator is simply to normalize the weights to unity. One can further normalize the weights to unity within subpopulations as defined by the covariates. In the limit this leads to an estimator proposed by Hirano, Imbens, and Ridder (2003), who suggest using a nonparametric series estimator for $e(x)$. More precisely, they first specify a sequence of functions of the covariates, such as power series $h_l(x)$, $l = 1, \ldots, \infty$. Next, they choose a number of terms, $L(N)$, as a function of the sample size, and then estimate the L-dimensional vector γ_L in

$$\Pr(W = 1|X = x) = \frac{\exp[(h_1(x), \ldots, h_L(x))\gamma_L]}{1 + \exp[(h_1(x), \ldots, h_L(x))\gamma_L]},$$

by maximizing the associated likelihood function. Let $\hat{\gamma}_L$ be the maximum likelihood estimate. In the third step, the estimated propensity score is calculated as

$$\hat{e}(x) = \frac{\exp[(h_1(x), \ldots, h_L(x))\hat{\gamma}_L]}{1 + \exp[(h_1(x), \ldots, h_L(x))\hat{\gamma}_L]}.$$

Finally they estimate the average treatment effect as

$$\hat{\tau}_{\text{weight}} = \sum_{i=1}^{N} \frac{W_i \cdot Y_i}{\hat{e}(X_i)} \Bigg/ \sum_{i=1}^{N} \frac{W_i}{\hat{e}(X_i)} - \sum_{i=1}^{N} \frac{(1 - W_i) \cdot Y_i}{1 - \hat{e}(X_i)} \Bigg/ \sum_{i=1}^{N} \frac{1 - W_i}{1 - \hat{e}(X_i)}. \tag{7}$$

Hirano, Imbens, and Ridder show that with a nonparametric estimator for $e(x)$ this estimator is efficient, whereas with the true propensity score the estimator would not be fully efficient (and in fact not very attractive).

This estimator highlights one of the interesting features of the problem of efficiently estimating average treatment effects. One solution is to estimate the two regression functions $\mu_w(x)$ nonparametrically, as discussed in Section IIIA; that solution completely ignores the propensity score. A second approach is to estimate the propensity score nonparametrically, ignoring entirely the two regression functions. If appropriately implemented, both approaches lead to fully efficient estimators, but clearly their finite-sample properties may be very different, depending, for example, on the smoothness of the regression functions versus the smoothness of the propensity score. If there is only a single binary covariate, or more generally if there are only discrete covariates, the weighting approach with a fully nonparametric estimator for the propensity score is numerically identical to the regression approach with a fully nonparametric estimator for the two regression functions.

To estimate the average treatment effect for the treated rather than for the full population, one should weight the contribution for unit i by the propensity score $e(x_i)$. If the propensity score is known, this leads to

$$\hat{\tau}_{\text{weight,tr}} = \sum_{i=1}^{N} W_i \cdot Y_i \cdot \frac{e(X_i)}{\hat{e}(X_i)} \Bigg/ \sum_{i=1}^{N} W_i \frac{e(X_i)}{\hat{e}(X_i)} - \sum_{i=1}^{N} (1 - W_i) \cdot Y_i \cdot \frac{e(X_i)}{1 - \hat{e}(X_i)} \Bigg/ \sum_{i=1}^{N} (1 - W_i) \frac{e(X_i)}{1 - \hat{e}(X_i)},$$

where the propensity score enters in some places as the true score (for the weights to get the appropriate estimand) and in other cases as the estimated score (to achieve efficiency). In the unknown propensity score case one always uses the estimated propensity score, leading to

$$\hat{\tau}_{\text{weight,tr}} = \left[\frac{1}{N_1} \sum_{i\,:\,W_i=1} Y_i\right] - \left[\sum_{i\,:\,W_i=0} Y_i \cdot \frac{\hat{e}(X_i)}{1 - \hat{e}(X_i)} \Bigg/ \sum_{i\,:\,W_i=0} \frac{\hat{e}(X_i)}{1 - \hat{e}(X_i)}\right].$$

One difficulty with the weighting estimators that are based on the estimated propensity score is again the problem of choosing the smoothing parameters. Hirano, Imbens, and Ridder (2003) use series estimators, which requires choosing the number of terms in the series. Ichimura and Linton (2001) consider a kernel version, which involves choosing a bandwidth. Theirs is currently one of the few studies considering optimal choices for smoothing parameters that focuses specifically on estimating average treatment effects. A departure from standard problems in choosing smoothing parameters is that here one wants to use nonparametric regression methods even if the propensity score is known. For example, if the probability of treatment is constant, standard optimality results would suggest using a high degree of smoothing, as this would lead to the most accurate estimator for the propensity score. However, this would not necessarily lead to an efficient estimator for the average treatment effect of interest.

Blocking on the Propensity Score: In their original propensity-score paper Rosenbaum and Rubin (1983a) suggest the following *blocking-on-the-propensity-score* estimator. Using the (estimated) propensity score, divide the sample into M blocks of units of approximately equal probability of treatment, letting J_{im} be an indicator for unit i being in block m. One way of implementing this is by dividing the unit interval into M blocks with boundary values equal to m/M for $m = 1, \ldots, M - 1$, so that

$$J_{im} = 1\left\{\frac{m - 1}{M} < e(X_i) \leq \frac{m}{M}\right\}$$

for $m = 1, \ldots, M$. Within each block there are N_{wm} observations with treatment equal to w, $N_{wm} = \sum_i 1\{W_i = w, J_{im} = 1\}$. Given these subgroups, estimate within each block the average treatment effect as if random assignment held:

$$\hat{\tau}_m = \frac{1}{N_{1m}} \sum_{i=1}^{N} J_{im} W_i Y_i - \frac{1}{N_{0m}} \sum_{i=1}^{N} J_{im}(1 - W_i) Y_i.$$

Then estimate the overall average treatment effect as

$$\hat{\tau}_{\text{block}} = \sum_{m=1}^{M} \hat{\tau}_m \cdot \frac{N_{1m} + N_{0m}}{N}.$$

If one is interested in the average effect for the treated, one will weight the within-block average treatment effects by the number of treated units:

$$\hat{\tau}_{T,\text{block}} = \sum_{m=1}^{M} \hat{\tau}_m \cdot \frac{N_{1m}}{N_T}.$$

Blocking can be interpreted as a crude form of nonparametric regression where the unknown function is approximated by a step function with fixed jump points. To establish asymptotic properties for this estimator would require establishing conditions on the rate at which the number of blocks increases with the sample size. With the propensity score known, these are easy to determine; no formal results have been established for the unknown propensity score case.

The question arises how many blocks to use in practice. Cochran (1968) analyzes a case with a single covariate and, assuming normality, shows that using five blocks removes at least 95% of the bias associated with that covariate. Since all bias, under unconfoundedness, is associated with the propensity score, this suggests that under normality the use of five blocks removes most of the bias associated with all the covariates. This has often been the starting point of empirical analyses using this estimator (for example, Rosenbaum and Rubin, 1983b; Dehejia and Wahba, 1999) and has been implemented in Stata by Becker and Ichino (2002).[7] Often, however, researchers subsequently check the balance of the covariates within each block. If the true propensity score per block is constant, the distribution of the covariates among the treated and controls should be identical, or, in the evaluation terminology, the covariates should be balanced. Hence one can assess the adequacy of the statistical model by comparing the distribution of the covariates among treated and controls within blocks. If the distributions are found to be different, one can either split the blocks into a number of subblocks, or generalize the specification of the propensity score. Often some informal version of the following algorithm is used: If within a block the propensity score itself is unbalanced, the blocks are too large and need to be split. If, conditional on the propensity score being balanced, the covariates are unbalanced, the specification of the propensity score is not adequate. No formal algorithm has been proposed for implementing these blocking methods.

An alternative approach to finding the optimal number of blocks is to relate this approach to the weighting estimator discussed above. One can view the blocking estimator as identical to a weighting estimator, with a modified estimator for the propensity score. Specifically, given the original estimator $\hat{e}(x)$, in the blocking approach the estimator for the propensity score is discretized to

$$\tilde{e}(x) = \frac{1}{M} \sum_{m=1}^{M} 1\left\{\frac{m}{M} \leq \hat{e}(x)\right\}.$$

Using $\tilde{e}(x)$ as the propensity score in the weighting estimator leads to an estimator for the average treatment effect identical to that obtained by using the blocking estimator with $\hat{e}(x)$ as the propensity score and M blocks. With sufficiently large M, the blocking estimator is sufficiently close to the original weighting estimator that it shares its first-order asymptotic properties, including its efficiency. This suggests that in general there is little harm in choosing a large number of blocks, at least with regard to asymptotic properties, although again the relevance of this for finite samples has not been established.

Regression on the Propensity Score: The third method of using the propensity score is to estimate the conditional expectation of Y given W and $e(X)$. Define

$$\nu_w(e) = \mathbb{E}[Y(w)|e(X) = e].$$

By unconfoundedness this is equal to $\mathbb{E}[Y|W = w, e(X) = e]$. Given an estimator $\hat{\nu}_w(e)$, one can estimate the average treatment effect as

$$\hat{\tau}_{\text{regprop}} = \frac{1}{N} \sum_{i=1}^{N} [\hat{\nu}_1(e(X_i)) - \hat{\nu}_0(e(X_i))].$$

Heckman, Ichimura, and Todd (1998) consider a local linear version of this for estimating the average treatment effect for the treated. Hahn (1998) considers a series version and shows that it is not as efficient as the regression estimator based on adjustment for all covariates.

Matching on the Propensity Score: Rosenbaum and Rubin's result implies that it is sufficient to adjust solely for differences in the propensity score between treated and control units. Since one of the ways in which one can adjust for differences in covariates is matching, another natural

[7] Becker and Ichino also implement estimators that match on the propensity score.

way to use the propensity score is through matching. Because the propensity score is a scalar function of the covariates, the bias results in Abadie and Imbens (2002) imply that the bias term is of lower order than the variance term and matching leads to a $\sqrt{N}$-consistent, asymptotically normally distributed estimator. The variance for the case with matching on the true propensity score also follows directly from their results. More complicated is the case with matching on the estimated propensity score. I do not know of any results that give the variance for this case.

D. Mixed Methods

A number of approaches have been proposed that combine two of the three methods described in the previous sections, typically regression with one of its alternatives. The reason for these combinations is that, although one method alone is often sufficient to obtain consistent or even efficient estimates, incorporating regression may eliminate remaining bias and improve precision. This is particularly useful in that neither matching nor the propensity-score methods directly address the correlation between the covariates and the outcome. The benefit associated with combining methods is made explicit in the notion developed by Robins and Ritov (1997) of *double robustness.* They propose a combination of weighting and regression where, as long as the parametric model for either the propensity score or the regression functions is specified correctly, the resulting estimator for the average treatment effect is consistent. Similarly, matching leads to consistency without additional assumptions; thus methods that combine matching and regressions are robust against misspecification of the regression function.

Weighting and Regression: One can rewrite the weighting estimator discussed above as estimating the following regression function by weighted least squares:

$$Y_i = \alpha + \tau \cdot W_i + \varepsilon_i,$$

with weights equal to

$$\lambda_i = \sqrt{\frac{W_i}{e(X_i)} + \frac{1 - W_i}{1 - e(X_i)}}.$$

Without the weights the least squares estimator would not be consistent for the average treatment effect; the weights ensure that the covariates are uncorrelated with the treatment indicator and hence the weighted estimator is consistent.

This weighted-least-squares representation suggests that one may add covariates to the regression function to improve precision, for example,

$$Y_i = \alpha + \beta' X_i + \tau \cdot W_i + \varepsilon_i,$$

with the same weights λ_i. Such an estimator, using a more general semiparametric regression model, was suggested by Robins and Rotnitzky (1995), Robins, Roznitzky, and Zhao (1995), and Robins and Ritov (1997), and implemented by Hirano and Imbens (2001). In the parametric context Robins and Ritov argue that the estimator is consistent as long as either the regression model or the propensity score (and thus the weights) are specified correctly. That is, in Robins and Ritov's terminology, the estimator is doubly robust.

Blocking and Regression: Rosenbaum and Rubin (1983b) suggest modifying the basic blocking estimator by using least squares regression within the blocks. Without the additional regression adjustment the estimated treatment effect within blocks can be written as a least squares estimator of τ_m for the regression function

$$Y_i = \alpha_m + \tau_m \cdot W_i + \varepsilon_i,$$

using only the units in block m. As above, one can also add covariates to the regression function

$$Y_i = \alpha_m + \beta_m' X_i + \tau_m \cdot W_i + \varepsilon_i,$$

again estimated on the units in block m.

Matching and Regression: Because Abadie and Imbens (2002) have shown that the bias of the simple matching estimator can dominate the variance if the dimension of the covariates is too large, additional bias corrections through regression can be particularly relevant in this case. A number of such corrections have been proposed, first by Rubin (1973b) and Quade (1982) in a parametric setting. Following the notation of section IIIB, let $\hat{Y}_i(0)$ and $\hat{Y}_i(1)$ be the observed or imputed potential outcomes for unit i; the estimated potential outcomes equal the observed outcomes for some unit i and for its match $\ell(i)$. The bias in their comparison, $\mathbb{E}[\hat{Y}_i(1) - \hat{Y}_i(0)] - [Y_i(1) - Y_i(0)]$, arises from the fact that the covariates X_i and $X_{\ell(i)}$ for units i and $\ell(i)$ are not equal, although they are close because of the matching process.

To further explore this, focusing on the single-match case, define for each unit

$$\hat{X}_i(0) = \begin{cases} X_i & \text{if } W_i = 0, \\ X_{\ell_1(i)} & \text{if } W_i = 1 \end{cases}$$

and

$$\hat{X}_i(1) = \begin{cases} X_{\ell_1(i)} & \text{if } W_i = 0, \\ X_i & \text{if } W_i = 1. \end{cases}$$

If the matching is exact, $\hat{X}_i(0) = \hat{X}_i(1)$ for each unit. If not, these discrepancies may lead to bias. The difference $\hat{X}_i(1) - \hat{X}_i(0)$ will therefore be used to reduce the bias of the simple matching estimator.

Suppose unit i is a treated unit ($W_i = 1$), so that $\hat{Y}_i(1) = Y_i(1)$ and $\hat{Y}_i(0)$ is an imputed value for $Y_i(0)$. This imputed value is unbiased for $\mu_0(X_{\ell_1(i)})$ (since $\hat{Y}_i(0) = Y_{\ell(i)}$), but not necessarily for $\mu_0(X_i)$. One may therefore wish to adjust $\hat{Y}_i(0)$ by an estimate of $\mu_0(X_i) - \mu_0(X_{\ell_1(i)})$. Typically these corrections are taken to be linear in the difference in the covariates for unit i and its match, that is, of the form $\beta_0'[\hat{X}_i(1) - \hat{X}_i(0)] = \beta_0'(X_i - X_{\ell_1(i)})$. Rubin (1973b) proposed three corrections, which differ in how β_0 is estimated.

To introduce Rubin's first correction, note that one can write the matching estimator as the least squares estimator for the regression function

$$\hat{Y}_i(1) - \hat{Y}_i(0) = \tau + \varepsilon_i.$$

This representation suggests modifying the regression function to

$$\hat{Y}_i(1) - \hat{Y}_i(0) = \tau + [\hat{X}_i(1) - \hat{X}_i(0)]'\beta + \varepsilon_i,$$

and again estimating τ by least squares.

The second correction is to estimate $\mu_0(x)$ directly by taking all control units, and estimate a linear regression of the form

$$Y_i = \alpha_0 + \beta_0' X_i + \varepsilon_i$$

by least squares. [If unit i is a control unit, the correction will be done using an estimator for the regression function $\mu_1(x)$ based on a linear specification $Y_i = \alpha_1 + \beta_1' X_i$ estimated on the treated units.] Abadie and Imbens (2002) show that if this correction is done nonparametrically, the resulting matching estimator is consistent and asymptotically normal, with its bias dominated by the variance.

The third method is to estimate the same regression function for the controls, but using only those that are used as matches for the treated units, with weights corresponding to the number of times a control observations is used as a match (see Abadie and Imbens, 2002). Compared to the second method, this approach may be less efficient, as it discards some control observations and weights some more than others. It has the advantage, however, of only using the most relevant matches. The controls that are discarded in the matching process are likely to be outliers relative to the treated observations, and they may therefore unduly affect the least squares estimates. If the regression function is in fact linear, this may be an attractive feature, but if there is uncertainty over its functional form, one may not wish to allow these observations such influence.

E. Bayesian Approaches

Little has been done using Bayesian methods to estimate average treatment effects, either in methodology or in application. Rubin (1978) introduces a general approach to estimating average and distributional treatment effects from a Bayesian perspective. Dehejia (2002) goes further, studying the policy decision problem of assigning heterogeneous individuals to various training programs with uncertain and variable effects.

To my knowledge, however, there are no applications using the Bayesian approach that focus on estimating the average treatment effect under unconfoundedness, either for the whole population or just for the treated. Neither are there simulation studies comparing operating characteristics of Bayesian methods with the frequentist methods discussed in the earlier sections of this paper. Such a Bayesian approach can be easily implemented with the regression methods discussed in section IIIA. Interestingly, it is less clear how Bayesian methods would be used with pairwise matching, which does not appear to have a natural likelihood interpretation.

A Bayesian approach to the regression estimators may be useful for a number of reasons. First, one of the leading problems with regression estimators is the presence of many covariates relative to the number of observations. Standard frequentist methods tend to either include those covariates without any restrictions, or exclude them entirely. In contrast, Bayesian methods would allow researchers to include covariates with more or less informative prior distributions. For example, if the researcher has a number of lagged outcomes, one may expect recent lags to be more important in predicting future outcomes than longer lags; this can be reflected in tighter prior distributions around zero for the older information. Alternatively, with a number of similar covariates one may wish to use hierarchical models that avoid problems with large-dimensional parameter spaces.

A second argument for considering Bayesian methods is that in an area closely related to this process of estimated unobserved outcomes—that of missing data with the missing at random (MAR) assumption—Bayesian methods have found widespread applicability. As advocated by Rubin (1987), multiple imputation methods often rely on a Bayesian approach for imputing the missing data, taking account of the parameter heterogeneity in a manner consistent with the uncertainty in the missing-data model itself. The same methods could be used with little modification for causal models, with the main complication that a relatively large proportion—namely 50% of the total number of potential outcomes—is missing.

IV. Estimating Variances

The variances of the estimators considered so far typically involve unknown functions. For example, as discussed in section IIE, the variance of efficient estimators of the PATE is equal to

$$V^P = \mathbb{E}\left[\frac{\sigma_1^2(X)}{e(X)} + \frac{\sigma_0^2(X)}{1 - e(X)} + (\mu_1(X) - \mu_0(X) - \tau)^2\right],$$

involving the two regression functions, the two conditional variances, and the propensity score.

There are a number of ways we can estimate this asymptotic variance. The first is essentially by brute force. All five components of the variance, $\sigma_0^2(x)$, $\sigma_1^2(x)$, $\mu_0(x)$, $\mu_1(x)$, and $e(x)$, are consistently estimable using kernel methods or series, and hence the asymptotic variance can be estimated consistently. However, if one estimates the average treatment effect using only the two regression functions, it is an additional burden to estimate the conditional variances and the propensity score in order to estimate V^P. Similarly, if one efficiently estimates the average treatment effect by weighting with the estimated propensity score, it is a considerable additional burden to estimate the first two moments of the conditional outcome distributions just to estimate the asymptotic variance.

A second method applies to the case where either the regression functions or the propensity score is estimated using series or sieves. In that case one can interpret the estimators, given the number of terms in the series, as parametric estimators, and calculate the variance this way. Under some conditions that will lead to valid standard errors and confidence intervals.

A third approach is to use bootstrapping (Efron and Tibshirani, 1993; Horowitz, 2002). There is little formal evidence specific for these estimators, but, given that the estimators are asymptotically linear, it is likely that bootstrapping will lead to valid standard errors and confidence intervals at least for the regression and propensity score methods. Bootstrapping may be more complicated for matching estimators, as the process introduces discreteness in the distribution that will lead to ties in the matching algorithm. Subsampling (Politis and Romano, 1999) will still work in this setting.

These first three methods provide variance estimates for estimators of τ^P. As argued above, however, one may instead wish to estimate τ^S or $\tau(X)$, in which case the appropriate (conservative) variance is

$$V^S = \mathbb{E}\left[\frac{\sigma_1^2(X)}{e(X)} + \frac{\sigma_0^2(X)}{1-e(X)}\right].$$

As above, this variance can be estimated by estimating the conditional moments of the outcome distributions, with the accompanying inherent difficulties. V^S cannot, however, be estimated by bootstrapping, since the estimand itself changes across bootstrap samples.

There is, however, an alternative method for estimating this variance that does not require additional nonparametric estimation. The idea behind this matching variance estimator, as developed by Abadie and Imbens (2002), is that even though the asymptotic variance depends on the conditional variance $\sigma_w^2(x)$, one need not actually estimate this variance consistently at all values of the covariates. Rather, one needs only the average of this variance over the distribution, weighted by the inverse of either $e(x)$ or its complement $1 - e(x)$. The key is therefore to obtain a close-to-unbiased estimator for the variance $\sigma_w^2(x)$. More generally, suppose we can find two treated units with $X = x$, say units i and j. In that case an unbiased estimator for $\sigma_1^2(x)$ is

$$\hat{\sigma}_1^2(x) = (Y_i - Y_j)^2/2.$$

In general it is again difficult to find exact matches, but again, this is not necessary. Instead, one uses the closest match within the set of units with the same treatment indicator. Let $v_m(i)$ be the mth closest unit to i with the same treatment indicator ($W_{v_m(i)} = W_i$), and

$$\sum_{l|W_l=W_i, l\neq i} 1\{\|X_l - x\| \leq \|X_{v_m(i)} - x\|\} = m.$$

Given a fixed number of matches, M, this gives us M units with the same treatment indicator and approximately the same values for the covariates. The sample variance of the outcome variable for these M units can then be used to estimate $\sigma_1^2(x)$. Doing the same for the control variance function, $\sigma_0^2(x)$, we can estimate $\sigma_w^2(x)$ at all values of the covariates and for $w = 0, 1$.

Note that these are not consistent estimators of the conditional variances. As the sample size increases, the bias of these estimators will disappear, just as we saw that the bias of the matching estimator for the average treatment effect disappears under similar conditions. The rate at which this bias disappears depends on the dimension of the covariates. The variance of the estimators for $\sigma_w^2(X_i)$, namely at specific values of the covariates, will not go to zero; however, this is not important, as we are interested not in the variances at specific points in the covariates distribution, but in the variance of the average treatment effect, V^S. Following the process introduce above, this last step is estimated as

$$\hat{V}^S = \frac{1}{N}\sum_{i=1}^{N}\left(\frac{\hat{\sigma}_1^2(X_i)}{\hat{e}(X_i)} + \frac{\hat{\sigma}_0^2(X_i)}{1-\hat{e}(X_i)}\right).$$

Under standard regularity conditions this is consistent for the asymptotic variance of the average treatment effect estimator. For matching estimators even estimation of the propensity score can be avoided. Abadie and Imbens show that one can estimate the variance of the matching estimator for SATE as:

$$\hat{V}^E = \frac{1}{N}\sum_{i=1}^{N}\left(1 + \frac{K_M(i)}{M}\right)^2 \hat{\sigma}_{W_i}^2(X_i),$$

where M is the number of matches and $K_M(i)$ is the number of times unit i is used as a match.

V. Assessing the Assumptions

A. *Indirect Tests of the Unconfoundedness Assumption*

The unconfoundedness assumption relied upon throughout this discussion is not directly testable. As discussed

above, it states that the conditional distribution of the outcome under the control treatment, $Y(0)$, given receipt of the active treatment and given covariates, is identical to the distribution of the control outcome given receipt of the control treatment and given covariates. The same is assumed for the distribution of the active treatment outcome, $Y(1)$. Because the data are completely uninformative about the distribution of $Y(0)$ for those who received the active treatment and of $Y(1)$ for those who received the control, the data cannot directly reject the unconfoundedness assumption. Nevertheless, there are often indirect ways of assessing this assumption, a number of which are developed in Heckman and Hotz (1989) and Rosenbaum (1987). These methods typically rely on estimating a causal effect that is known to equal zero. If the test then suggests that this causal effect differs from zero, the unconfoundedness assumption is considered less plausible. These tests can be divided into two broad groups.

The first set of tests focuses on estimating the causal effect of a treatment that is known not to have an effect, relying on the presence of multiple control groups (Rosenbaum, 1987). Suppose one has two potential control groups, for example, eligible nonparticipants and ineligibles, as in Heckman, Ichimura, and Todd (1997). One interpretation of the test is to compare average treatment effects estimated using each of the control groups. This can also be interpreted as estimating an "average treatment effect" using only the two control groups, with the treatment indicator now a dummy for being a member of the first group. In that case the treatment effect is known to be zero, and statistical evidence of a nonzero effect implies that at least one of the control groups is invalid. Again, not rejecting the test does not imply the unconfoundedness assumption is valid (as both control groups could suffer the same bias), but nonrejection in the case where the two control groups are likely to have different potential biases makes it more plausible that the unconfoundedness assumption holds. The key for the power of this test is to have available control groups that are likely to have different biases, if any. Comparing ineligibles and eligible nonparticipants as in Heckman, Ichimura, and Todd (1997) is a particularly attractive comparison. Alternatively one may use different geographic controls, for example from areas bordering on different sides of the treatment group.

One can formalize this test by postulating a three-valued indicator $T_i \in \{-1, 0, 1\}$ for the groups (e.g., ineligibles, eligible nonparticipants, and participants), with the treatment indicator equal to $W_i = 1\{T_i = 1\}$. If one extends the unconfoundedness assumption to independence of the potential outcomes and the group indicator given covariates,

$$Y_i(0),\ Y_i(1) \perp T_i | X_i,$$

then a testable implication is

$$Y_i \perp 1\{T_i = 0\} | X_i,\ T_i \leq 0.$$

An implication of this independence condition is being tested by the tests discussed above. Whether this test has much bearing on the unconfoundedness assumption depends on whether the extension of the assumption is plausible given unconfoundedness itself.

The second set of tests of unconfoundedness focuses on estimating the causal effect of the treatment on a variable known to be unaffected by it, typically because its value is determined prior to the treatment itself. Such a variable can be time-invariant, but the most interesting case is in considering the treatment effect on a lagged outcome. If this is not zero, this implies that the treated observations are distinct from the controls; namely, that the distribution of $Y_{i,-1}$ for the treated units is not comparable to the distribution of $Y_{i,-1}$ for the controls. If the treatment is instead zero, it is more plausible that the unconfoundedness assumption holds. Of course this does not directly test this assumption; in this setting, being able to reject the null of no effect does not directly reflect on the hypothesis of interest, unconfoundedness. Nevertheless, if the variables used in this proxy test are closely related to the outcome of interest, the test arguably has more power. For these tests it is clearly helpful to have a number of lagged outcomes.

To formalize this, let us suppose the covariates consist of a number of lagged outcomes $Y_{i,-1}, \ldots, Y_{i,-T}$ as well as time-invariant individual characteristics Z_i, so that $X_i = (Y_{i,-1}, \ldots, Y_{i,-T}, Z_i)$. By construction only units in the treatment group after period -1 receive the treatment; all other observed outcomes are control outcomes. Also suppose that the two potential outcomes $Y_i(0)$ and $Y_i(1)$ correspond to outcomes in period zero. Now consider the following two assumptions. The first is unconfoundedness given only $T - 1$ lags of the outcome:

$$Y_i(1),\ Y_i(0) \perp W_i | Y_{i,-1}, \ldots, Y_{i,-(T-1)}, Z_i,$$

and the second assumes stationarity and exchangeability:

$$f_{Y_{i,s}(0)|Y_{i,s-1}(0),\ldots,Y_{i,s-(T-1)}(0),Z_i,W_i}(y_s | y_{s-1}, \ldots, y_{s-(T-1)}, z, w)$$

does not depend on i and s. Then it follows that

$$Y_{i,-1} \perp W_i | Y_{i,-2}, \ldots, Y_{i,-T}, Z_i,$$

which is testable. This hypothesis is what the test described above tests. Whether this test has much bearing on unconfoundedness depends on the link between the two assumptions and the original unconfoundedness assumption. With a sufficient number of lags, unconfoundedness given all lags but one appears plausible, conditional on unconfoundedness given all lags, so the relevance of the test depends largely on the plausibility of the second assumption, stationarity and exchangeability.

B. *Choosing the Covariates*

The discussion so far has focused on the case where the covariates set is known a priori. In practice there can be two

issues with the choice of covariates. First, there may be some variables that should not be adjusted for. Second, even with variables that should be adjusted for in large samples, the expected mean squared error may be reduced by ignoring those covariates that have only weak correlation with the treatment indicator and the outcomes. This second issue is essentially a statistical one. Including a covariate in the adjustment procedure, through regression, matching or otherwise, will not lower the asymptotic precision of the average treatment effect if the assumptions are correct. In finite samples, however, a covariate that is not, or is only weakly, correlated with outcomes and treatment indicators may reduce precision. There are few procedures currently available for optimally choosing the set of covariates to be included in matching or regression adjustments, taking into account such finite-sample properties.

The first issue is a substantive one. The unconfoundedness assumption may apply with one set of covariates but not apply with an expanded set. A particular concern is the inclusion of covariates that are themselves affected by the treatment, such as intermediate outcomes. Suppose, for example, that in evaluating a job training program, the primary outcome of interest is earnings two years later. In that case, employment status prior to the program is unaffected by the treatment and thus a valid element of the set of adjustment covariates. In contrast, employment status one year after the program is an intermediate outcome and should not be controlled for. It could itself be an outcome of interest, and should therefore never be a covariate in an analysis of the effect of the training program. One guarantee that a covariate is not affected by the treatment is that it was measured before the treatment was chosen. In practice, however, the covariates are often recorded at the same time as the outcomes, subsequent to treatment. In that case one has to assess on a case-by-case basis whether a particular covariate should be used in adjusting outcomes. See Rosenbaum (1984b) and Angrist and Krueger (2000) for more discussion.

C. *Assessing the Overlap Assumption*

The second of the key assumptions in estimating average treatment effects requires that the propensity score—the probability of receiving the active treatment—be strictly between zero and one. In principle this is testable, as it restricts the joint distribution of observables; but formal tests are not necessarily the main concern. In practice, this assumption raises a number of questions. The first is how to detect a lack of overlap in the covariate distributions. A second is how to deal with it, given that such a lack exists. A third is how the individual methods discussed in section III address this lack of overlap. Ideally such a lack would result in large standard errors for the average treatment effects.

The first method to detect lack of overlap is to plot distributions of covariates by treatment groups. In the case with one or two covariates one can do this directly. In high-dimensional cases, however, this becomes more difficult. One can inspect pairs of marginal distributions by treatment status, but these are not necessarily informative about lack of overlap. It is possible that for each covariate the distributions for the treatment and control groups are identical, even though there are areas where the propensity score is 0 or 1.

A more useful method is therefore to inspect the distribution of the propensity score in both treatment groups, which can directly reveal lack of overlap in high-dimensional covariate distributions. Its implementation requires nonparametric estimation of the propensity score, however, and misspecification may lead to failure in detecting a lack of overlap, just as inspecting various marginal distributions may be insufficient. In practice one may wish to undersmooth the estimation of the propensity score, either by choosing a bandwidth smaller than optimal for nonparametric estimation or by including higher-order terms in a series expansion.

A third way to detect lack of overlap is to inspect the quality of the worst matches in a matching procedure. Given a set of matches, one can, for each component k of the vector of covariates, inspect $\max_i |x_{i,k} - x_{\ell_1(i),k}|$, the maximum over all observations of the matching discrepancy. If this difference is large relative to the sample standard deviation of the kth component of the covariates, there is reason for concern. The advantage of this method is that it does not require additional nonparametric estimation.

Once one determines that there is a lack of overlap, one can either conclude that the average treatment effect of interest cannot be estimated with sufficient precision, and/or decide to focus on an average treatment effect that is estimable with greater accuracy. To do the latter it can be useful to discard some of the observations on the basis of their covariates. For example, one may decide to discard control (treated) observations with propensity scores below (above) a cutoff level. The desired cutoff may depend on the sample size; in a very large sample one may not be concerned with a propensity score of 0.01, whereas in small samples such a value may make it difficult to find reasonable comparisons. To judge such tradeoffs, it is useful to understand the relationship between a unit's propensity score and its implicit weight in the average-treatment-effect estimation. Using the weighting estimator, the average outcome under the treatment is estimated by summing up outcomes for the control units with weight approximately equal to 1 divided by their propensity score (and 1 divided by 1 minus the propensity score for treated units). Hence with N units, the weight of unit i is approximately $1/\{N \cdot [1 - e(X_i)]\}$ if it is a treated unit and $1/[N \cdot e(X_i)]$ if it is a control. One may wish to limit this weight to some fraction, for example, 0.05, so that no unit will have a weight of more than 5% in the average. Under that approach, the limit on the propensity score in a sample with

200 units is 0.1; units with a propensity score less than 0.1 or greater than 0.9 should be discarded. In a sample with 1000 units, only units with a propensity score outside the range [0.02, 0.98] will be ignored.

In matching procedures one need not rely entirely on comparisons of the propensity score distribution in discarding the observations with insufficient match quality. Whereas Rosenbaum and Rubin (1984) suggest accepting only matches where the difference in propensity scores is below a cutoff point, alternatively one may wish to drop matches where individual covariates are severely mismatched.

Finally, let us consider the three approaches to inference—regression, matching, and propensity score methods—and assess how each handles lack of overlap. Suppose one is interested in estimating the average effect on the treated, and one has a data set with sufficient overlap. Now suppose one adds a few treated or control observations with covariate values rarely seen in the alternative treatment group. Adding treated observations with outlying values implies one cannot estimate the average treatment effect for the treated very precisely, because one lacks suitable controls against which to compare these additional units. Thus with methods appropriately dealing with limited overlap one will see the variance estimates increase. In contrast, adding control observations with outlying covariate values should have little effect, since such controls are irrelevant for the average treatment effect for the treated. Therefore, methods appropriately dealing with limited overlap should in this case show estimates approximately unchanged in bias and precision.

Consider first the regression approach. Conditional on a particular parametric specification for the regression function, adding observations with outlying values of the regressors leads to considerably more precise parameter estimates; such observations are influential precisely because of their outlying values. If the added observations are treated units, the precision of the estimated control regression function at these outlying values will be lower (since few if any control units are found in that region); thus the variance will increase, as it should. One should note, however, that the estimates in this region may be sensitive to the specification chosen. In contrast, by the nature of regression functions, adding control observations with outlying values will lead to a spurious increase in precision of the control regression function. Regression methods can therefore be misleading in cases with limited overlap.

Next, consider matching. In estimating the average treatment effect for the treated, adding control observations with outlying covariate values will likely have little affect on the results, since such observations are unlikely to be used as matches. The results would, however, be sensitive to adding treated observations with outlying covariate values, because these observations would be matched to inappropriate controls, leading to possibly biased estimates. The standard errors would largely be unaffected.

Finally, consider propensity-score estimates. Estimates of the probability of receiving treatment now include values close to 0 and 1. The values close to 0 for the control observations would cause little difficulty because these units would get close to zero weight in the estimation. The control observations with a propensity score close to 1, however, would receive high weights, leading to an increase in the variance of the average-treatment-effect estimator, correctly implying that one cannot estimate the average treatment effect very precisely. Blocking on the propensity score would lead to similar conclusions.

Overall, propensity score and matching methods (and likewise kernel-based regression methods) are better designed to cope with limited overlap in the covariate distributions than are parametric or semiparametric (series) regression models. In all cases it is useful to inspect histograms of the estimated propensity score in both groups to assess whether limited overlap is an issue.

VI. Applications

There are many studies using some form of unconfoundedness or selection on observables, ranging from simple least squares analyses to matching on the propensity score (for example, Ashenfelter and Card, 1985; LaLonde, 1986; Card and Sullivan, 1988; Heckman, Ichimura, and Todd, 1997; Angrist, 1998; Dehejia and Wahba, 1999; Lechner, 1998; Friedlander and Robins, 1995; and many others). Here I focus primarily on two sets of analyses that can help researchers assess the value of the methods surveyed in this paper: first, studies attempting to assess the plausibility of the assumptions, often using randomized experiments as a yardstick; second, simulation studies focusing on the performance of the various techniques in settings where the assumptions are known to hold.

A. Applications: Randomized Experiments as Checks on Unconfoundedness

The basic idea behind these studies is simple: to use experimental results as a check on the attempted nonexperimental estimates. Given a randomized experiment, one can obtain unbiased estimates of the average effect of a program. Then, one can put aside the experimental control group and attempt to replicate these results using a nonexperimental control. If one can successfully replicate the experimental results, this suggests that the assumptions and methods are plausible. Such investigations are of course not generally conclusive, but are invaluable in assessing the plausibility of the approach. The first such study, and one that made an enormous impact in the econometrics literature, was by LaLonde (1986). Fraker and Maynard (1987) conducted a similar investigation, and many more have followed.

LaLonde (1986) took the National Supported Work program, a fairly small program aimed at particularly disadvantaged people in the labor market (individuals with poor labor market histories and skills). Using these data, he set aside the experimental control group and in its place constructed alternative controls from the Panel Study of Income Dynamics (PSID) and Current Population Survey (CPS), using various selection criteria depending on prior labor market experience. He then used a number of methods—ranging from a simple difference, to least squares adjustment, a Heckman selection correction, and difference-indifferences techniques—to create nonexperimental estimates of the average treatment effect. His general conclusion was that the results were very unpredictable and that no method could consistently replicate the experimental results using any of the six nonexperimental control groups constructed. A number of researchers have subsequently tested new techniques using these same data. Heckman and Hotz (1989) focused on testing the various models and argued that the testing procedures they developed would have eliminated many of LaLonde's particularly inappropriate estimates. Dehejia and Wahba (1999) used several of the semiparametric methods based on the unconfoundedness assumption discussed in this survey, and found that for the subsample of the LaLonde data that they used (with two years of prior earnings), these methods replicated the experimental results more accurately—both overall and within subpopulations. Smith and Todd (2003) analyze the same data and conclude that for other subsamples, including those for which only one year of prior earnings is available, the results are less robust. See Dehejia (2003) for additional discussion of these results.

Others have used different experiments to carry out the same or similar analyses, using varying sets of estimators and alternative control groups. Friedlander and Robins (1995) focus on least squares adjustment, using data from the WIN (Work INcentive) demonstration programs conducted in a number of states, and construct control groups from other counties in the same state, as well as from different states. They conclude that nonexperimental methods are unable to replicate the experimental results. Hotz, Imbens, and Mortimer (2003) use the same data and consider matching methods with various sets of covariates, using single or multiple alternative states as nonexperimental control groups. They find that for the subsample of individuals with positive earnings at some date prior to the program, nonexperimental methods work better than for those with no known positive earnings.

Heckman, Ichimura, and Todd (1997, 1998) and Heckman, Ichimura, Smith, and Todd (1998) study the national Job Training Partnership Act (JPTA) program, using data from different geographical locations to investigate the nature of the biases associated with different estimators, and the importance of overlap in the covariates, including labor market histories. Their conclusions provide the type of specific guidance that should be the aim of such studies. They give clear and generalizable conditions that make the assumptions of unconfoundedness and overlap—at least according to their study of a large training program—more plausible. These conditions include the presence of detailed earnings histories, and control groups that are geographically close to the treatment group—preferably groups of ineligibles, or eligible nonparticipants from the same location. In contrast, control groups from very different locations are found to be poor nonexperimental controls. Although such conclusions are only clearly generalizable to evaluations of social programs, they are potentially very useful in providing analysts with concrete guidance as to the applicability of these assumptions.

Dehejia (2002) uses the Greater Avenues to INdependence (GAIN) data, using different counties as well as different offices within the same county as nonexperimental control groups. Similarly, Hotz, Imbens, and Klerman (2001) use the basic GAIN data set supplemented with administrative data on long-term quarterly earnings (both prior and subsequent to the randomization date), to investigate the importance of detailed earnings histories. Such detailed histories can also provide more evidence on the plausibility of nonexperimental evaluations for long-term outcomes.

Two complications make this literature difficult to evaluate. One is the differences in covariates used; it is rare that variables are measured consistently across different studies. For instance, some have yearly earnings data, others quarterly, others only earnings indicators on a monthly or quarterly basis. This makes it difficult to consistently investigate the level of detail in earnings history necessary for the unconfoundedness assumption to hold. A second complication is that different estimators are generally used; thus any differences in results can be attributed to either estimators or assumptions. This is likely driven by the fact that few of the estimators have been sufficiently standardized that they can be implemented easily by empirical researchers.

All of these studies just discussed took data from actual randomized experiments to test the "true" treatment effect against the estimators used on the nonexperimental data. To some extent, however, such experimental data are not required. The question of interest is whether an alternative control group is an adequate proxy for a randomized control in a particular setting; note that this question does not require data on the treatment group. Although these questions have typically been studied by comparing experimental with nonexperimental results, all that is really relevant is whether the nonexperimental control group can predict the average outcomes for the experimental control. As in Heckman, Ichimura, Smith, and Todd's (1998) analysis of the JTPA data, one can take two groups, neither subject to the treatment, and ask the question whether—using data on the covariates for the first control group in combination with outcome and covariate information for the second—one can

predict the average outcome in the first. If so, this implies that, had there been an experiment on the population from which the first control group was drawn, the second group would provide an acceptable nonexperimental control. From this perspective one can use data from many different surveys. In particular, one can more systematically investigate whether control groups from different counties, states, or regions or even different time periods make acceptable nonexperimental controls.

B. Simulations

A second question that is often confounded with that of the validity of the assumptions is that of the relative performance of the various estimators. Suppose one is willing to accept the unconfoundedness and overlap assumptions. Which estimation method is most appropriate in a particular setting? In many of the studies comparing nonexperimental with experimental outcomes, researchers compare results for a number of the techniques described here. Yet in these settings we cannot be certain that the underlying assumptions hold. Thus, although it is useful to compare these techniques in such realistic settings, it is also important to compare them in an artificial environment where one is certain that the underlying assumptions are valid.

There exist a few studies that specifically set out to do this. Frölich (2000) compares a number of matching estimators and local linear regression methods, carefully formalizing fully data-driven procedures for the estimators considered. To make these comparisons he considers a large number of data-generating processes, based on eight different regression functions (including some highly nonlinear and multimodal ones), two different sample sizes, and three different density functions for the covariate (one important limitation is that he restricts the investigation to a single covariate). For the matching estimator Frölich considered a single match with replacement; for the local linear regression estimators he uses data-driven optimal bandwidth choices based on minimizing the mean squared error of the average treatment effect. The first local linear estimator considered is the standard one: at x the regression function $\mu(x)$ is estimated as β_0 in the minimization problem

$$\min_{\beta_0,\beta_1} \sum_{i=1}^{N} [Y_i - \beta_0 - \beta_1 \cdot (X_i - x)]^2 \cdot K\left(\frac{X_i - x}{h}\right),$$

with an Epanechnikov kernel. He finds that this has computational problems, as well as poor small-sample properties. He therefore also considers a modification suggested by Seifert and Gasser (1996, 2000). For given x, define $\bar{x} = \sum X_i K((X_i - x)/h)/\sum K((X_i - x)/h)$, so that one can write the standard local linear estimator as

$$\hat{\mu}(x) = \frac{T_0}{S_0} + \frac{T_1}{S_2}(x - \bar{x}),$$

where, for $r = 0, 1, 2$, one has $S_r = \sum K((X_i - x)/h)(X_i - x)^r$ and $T_r = \sum K((X_i - x)/h)(X_i - x)^r Y_i$. The Seifert-Gasser modification is to use instead

$$\hat{\mu}(x) = \frac{T_0}{S_0} + \frac{T_1}{S_2 + R}(x - \bar{x}),$$

where the recommended ridge parameter is $R = |x - \bar{x}|[5/(16h)]$, given the Epanechnikov kernel $k(u) = \frac{3}{4}(1 - u^2)1\{|u| < 1\}$. Note that with high-dimensional covariates, such a nonnegative kernel would lead to biases that do not vanish fast enough to be dominated by the variance (see the discussion in Heckman, Ichimura, and Todd, 1998). This is not a problem in Frölich's simulations, as he considers only cases with a single covariate. Frölich finds that the local linear estimator, with Seifert and Gassert's modification, performs better than either the matching or the standard local linear estimator.

Zhao (2004) uses simulation methods to compare matching and parametric regression estimators. He uses metrics based on the propensity score, the covariates, and estimated regression functions. Using designs with varying numbers of covariates and linear regression functions, Zhao finds there is no clear winner among the different estimators, although he notes that using the outcome data in choosing the metric appears a promising strategy.

Abadie and Imbens (2002) study their matching estimator using a data-generating process inspired by the LaLonde study to allow for substantial nonlinearity, fitting a separate binary response model to the zeros in the earnings outcome, and a log linear model for the positive observations. The regression estimators include linear and quadratic models (the latter with a full set of interactions), with seven covariates. This study finds that the matching estimators, and in particular the bias-adjusted alternatives, outperform the linear and quadratic regression estimators (the former using 7 covariates, the latter 35, after dropping squares and interactions that lead to perfect collinearity). Their simulations also suggest that with few matches—between one and four—matching estimators are not sensitive to the number of matches used, and that their confidence intervals have actual coverage rates close to the nominal values.

The results from these simulation studies are overall somewhat inconclusive; it is clear that more work is required. Future simulations may usefully focus on some of the following issues. First, it is obviously important to closely model the data-generating process on actual data sets, to ensure that the results have some relevance for practice. Ideally one would build the simulations around a number of specific data sets through a range of data-generating processes. Second, it is important to have fully data-driven procedures that define an estimator as a function of $(Y_i, W_i, X_i)_{i=1}^N$, as seen in Frölich (2000). For the matching estimators this is relatively straightforward, but for some others this requires more care. This will allow

other researchers to consider meaningful comparisons across the various estimators.

Finally, we need to learn which features of the data-generating process are important for the properties of the various estimators. For example, do some estimators deteriorate more rapidly than others when a data set has many covariates and few observations? Are some estimators more robust against high correlations between covariates and outcomes, or high correlations between covariates and treatment indicators? Which estimators are more likely to give conservative answers in terms of precision? Since it is clear that no estimator is always going to dominate all others, what is important is to isolate salient features of the data-generating processes that lead to preferring one alternative over another. Ideally we need descriptive statistics summarizing the features of the data that provide guidance in choosing the estimator that will perform best in a given situation.

VII. Conclusion

In this paper I have attempted to review the current state of the literature on inference for average treatment effects under the assumption of unconfoundedness. This has recently been a very active area of research where many new semi- and nonparametric econometric methods have been applied and developed. The research has moved a long way from relying on simple least squares methods for estimating average treatment effects.

The primary estimators in the current literature include propensity-score methods and pairwise matching, as well as nonparametric regression methods. Efficiency bounds have been established for a number of the average treatment effects estimable with these methods, and a variety of these estimators rely on the weakest assumptions that allow point identification. Researchers have suggested several ways for estimating the variance of these average-treatment-effect estimators. One, more cumbersome approach requires estimating each component of the variance nonparametrically. A more common method relies on bootstrapping. A third alternative, developed by Abadie and Imbens (2002) for the matching estimator, requires no additional nonparametric estimation. There is, as yet, however, no consensus on which are the best estimation methods to apply in practice. Nevertheless, the applied researcher has now a large number of new estimators at her disposal.

Challenges remain in making the new tools more easily applicable. Although software is available to implement some of the estimators (see Becker and Ichino, 2002; Sianesi, 2001; Abadie et al., 2003), many remain difficult to apply. A particularly urgent task is therefore to provide fully implementable versions of the various estimators that do not require the applied researcher to choose bandwidths or other smoothing parameters. This is less of a concern for matching methods and probably explains a large part of their popularity. Another outstanding question is the relative performance of these methods in realistic settings with large numbers of covariates and varying degrees of smoothness in the conditional means of the potential outcomes and the propensity score.

Once these issues have been resolved, today's applied evaluators will benefit from a new set of reliable, econometrically defensible, and robust methods for estimating the average treatment effect of current social policy programs under exogeneity assumptions.

REFERENCES

Abadie, A., "Semiparametric Instrumental Variable Estimation of Treatment Response Models," *Journal of Econometrics* 113:2 (2003a), 231–263.

Abadie, A., "Semiparametric Difference-in-Differences Estimators," forthcoming, *Review of Economic Studies* (2003b).

Abadie, A., J. Angrist, and G. Imbens, "Instrumental Variables Estimation of Quantile Treatment Effects," *Econometrica* 70:1 (2002), 91–117.

Abadie, A., D. Drukker, H. Herr, and G. Imbens, "Implementing Matching Estimators for Average Treatment Effects in STATA," Department of Economics, University of California, Berkeley, unpublished manuscript (2003).

Abadie, A., and G. Imbens, "Simple and Bias-Corrected Matching Estimators for Average Treatment Effects," NBER technical working paper no. 283 (2002).

Abbring, J., and G. van den Berg, "The Non-parametric Identification of Treatment Effects in Duration Models," Free University of Amsterdam, unpublished manuscript (2002).

Angrist, J., "Estimating the Labor Market Impact of Voluntary Military Service Using Social Security Data on Military Applicants," *Econometrica* 66:2 (1998), 249–288.

Angrist, J. D., and J. Hahn, "When to Control for Covariates? Panel-Asymptotic Results for Estimates of Treatment Effects," NBER technical working paper no. 241 (1999).

Angrist, J. D., G. W. Imbens, and D. B. Rubin, "Identification of Causal Effects Using Instrumental Variables," *Journal of the American Statistical Association* 91 (1996), 444–472.

Angrist, J. D., and A. B. Krueger, "Empirical Strategies in Labor Economics," in A. Ashenfelter and D. Card (Eds.), *Handbook of Labor Economics* vol. 3 (New York: Elsevier Science, 2000).

Angrist, J., and V. Lavy, "Using Maimonides' Rule to Estimate the Effect of Class Size on Scholastic Achievement," *Quarterly Journal of Economics* CXIV (1999), 1243.

Ashenfelter, O., "Estimating the Effect of Training Programs on Earnings," this REVIEW 60 (1978), 47–57.

Ashenfelter, O., and D. Card, "Using the Longitudinal Structure of Earnings to Estimate the Effect of Training Programs," this REVIEW 67 (1985), 648–660.

Athey, S., and G. Imbens, "Identification and Inference in Nonlinear Difference-in-Differences Models," NBER technical working paper no. 280 (2002).

Athey, S., and S. Stern, "An Empirical Framework for Testing Theories about Complementarity in Organizational Design," NBER working paper no. 6600 (1998).

Barnow, B. S., G. G. Cain, and A. S. Goldberger, "Issues in the Analysis of Selectivity Bias," in E. Stromsdorfer and G. Farkas (Eds.), *Evaluation Studies* vol. 5 (San Francisco: Sage, 1980).

Becker, S., and A. Ichino, "Estimation of Average Treatment Effects Based on Propensity Scores," *The Stata Journal* 2:4 (2002), 358–377.

Bitler, M., J. Gelbach, and H. Hoynes, "What Mean Impacts Miss: Distributional Effects of Welfare Reform Experiments," Department of Economics, University of Maryland, unpublished paper (2002).

Björklund, A., and R. Moffit, "The Estimation of Wage Gains and Welfare Gains in Self-Selection Models," this REVIEW 69 (1987), 42–49.

Black, S., "Do Better Schools Matter? Parental Valuation of Elementary Education," *Quarterly Journal of Economics* CXIV (1999), 577.

Blundell, R., and Monica Costa-Dias, "Alternative Approaches to Evaluation in Empirical Microeconomics," Institute for Fiscal Studies, Cemmap working paper cwp10/02 (2002).
Blundell, R., A. Gosling, H. Ichimura, and C. Meghir, "Changes in the Distribution of Male and Female Wages Accounting for the Employment Composition," Institute for Fiscal Studies, London, unpublished paper (2002).
Card, D., and D. Sullivan, "Measuring the Effect of Subsidized Training Programs on Movements In and Out of Employment," *Econometrica* 56:3 (1988), 497–530.
Chernozhukov, V., and C. Hansen, "An IV Model of Quantile Treatment Effects," Department of Economics, MIT, unpublished working paper (2001).
Cochran, W., "The Effectiveness of Adjustment by Subclassification in Removing Bias in Observational Studies," *Biometrics* 24, (1968), 295–314.
Cochran, W., and D. Rubin, "Controlling Bias in Observational Studies: A Review," *Sankhyā* 35 (1973), 417–446.
Dehejia, R., "Was There a Riverside Miracle? A Hierarchical Framework for Evaluating Programs with Grouped Data," *Journal of Business and Economic Statistics* 21:1 (2002), 1–11.
——— "Practical Propensity Score Matching: A Reply to Smith and Todd," forthcoming, *Journal of Econometrics* (2003).
Dehejia, R., and S. Wahba, "Causal Effects in Nonexperimental Studies: Reevaluating the Evaluation of Training Programs," *Journal of the American Statistical Association* 94 (1999), 1053–1062.
Doksum, K., "Empirical Probability Plots and Statistical Inference for Nonlinear Models in the Two-Sample Case," *Annals of Statistics* 2 (1974), 267–277.
Efron, B., and R. Tibshirani, *An Introduction to the Bootstrap* (New York: Chapman and Hall, 1993).
Engle, R., D. Hendry, and J.-F. Richard, "Exogeneity," *Econometrica* 51:2 (1974), 277–304.
Firpo, S., "Efficient Semiparametric Estimation of Quantile Treatment Effects," Department of Economics, University of California, Berkeley, PhD thesis (2002), chapter 2.
Fisher, R. A., *The Design of Experiments* (Boyd, London, 1935).
Fitzgerald, J., P. Gottschalk, and R. Moffitt, "An Analysis of Sample Attrition in Panel Data: The Michigan Panel Study of Income Dynamics," *Journal of Human Resources* 33 (1998), 251–299.
Fraker, T., and R. Maynard, "The Adequacy of Comparison Group Designs for Evaluations of Employment-Related Programs," *Journal of Human Resources* 22:2 (1987), 194–227.
Friedlander, D., and P. Robins, "Evaluating Program Evaluations: New Evidence on Commonly Used Nonexperimental Methods," *American Economic Review* 85 (1995), 923–937.
Frölich, M., "Treatment Evaluation: Matching versus Local Polynomial Regression," Department of Economics, University of St. Gallen, discussion paper no. 2000-17 (2000).
——— "What is the Value of Knowing the Propensity Score for Estimating Average Treatment Effects," Department of Economics, University of St. Gallen (2002).
Gill, R., and J. Robins, "Causal Inference for Complex Longitudinal Data: The Continuous Case," *Annals of Statistics* 29:6 (2001), 1785–1811.
Gu, X., and P. Rosenbaum, "Comparison of Multivariate Matching Methods: Structures, Distances and Algorithms," *Journal of Computational and Graphical Statistics* 2 (1993), 405–420.
Hahn, J., "On the Role of the Propensity Score in Efficient Semiparametric Estimation of Average Treatment Effects," *Econometrica* 66:2 (1998), 315–331.
Hahn, J., P. Todd, and W. Van der Klaauw, "Identification and Estimation of Treatment Effects with a Regression-Discontinuity Design," *Econometrica* 69:1 (2000), 201–209.
Ham, J., and R. LaLonde, "The Effect of Sample Selection and Initial Conditions in Duration Models: Evidence from Experimental Data on Training," *Econometrica* 64:1 (1996).
Heckman, J., and J. Hotz, "Alternative Methods for Evaluating the Impact of Training Programs" (with discussion), *Journal of the American Statistical Association* 84:804 (1989), 862–874.
Heckman, J., H. Ichimura, and P. Todd, "Matching as an Econometric Evaluation Estimator: Evidence from Evaluating a Job Training Program," *Review of Economic Studies* 64 (1997), 605–654.
——— "Matching as an Econometric Evaluation Estimator," *Review of Economic Studies* 65 (1998), 261–294.
Heckman, J., H. Ichimura, J. Smith, and P. Todd, "Characterizing Selection Bias Using Experimental Data," *Econometrica* 66 (1998), 1017–1098.
Heckman, J., R. LaLonde, and J. Smith, "The Economics and Econometrics of Active Labor Markets Programs," in A. Ashenfelter and D. Card (Eds.), *Handbook of Labor Economics* vol. 3 (New York: Elsevier Science, 2000).
Heckman, J., and R. Robb, "Alternative Methods for Evaluating the Impact of Interventions," in J. Heckman and B. Singer (Eds.), *Longitudinal Analysis of Labor Market Data* (Cambridge, U.K.: Cambridge University Press, 1984).
Heckman, J., J. Smith, and N. Clements, "Making the Most out of Programme Evaluations and Social Experiments: Accounting for Heterogeneity in Programme Impacts," *Review of Economic Studies* 64 (1997), 487–535.
Hirano, K., and G. Imbens, "Estimation of Causal Effects Using Propensity Score Weighting: An Application of Data on Right Ear Catheterization," *Health Services and Outcomes Research Methodology* 2 (2001), 259–278.
Hirano, K., G. Imbens, and G. Ridder, "Efficient Estimation of Average Treatment Effects Using the Estimated Propensity Score," *Econometrica* 71:4 (2003), 1161–1189.
Holland, P., "Statistics and Causal Inference" (with discussion), *Journal of the American Statistical Association* 81 (1986), 945–970.
Horowitz, J., "The Bootstrap," in James J. Heckman and E. Leamer (Eds.), *Handbook of Econometrics,* vol. 5 (Elsevier North Holland, 2002).
Hotz, J., G. Imbens, and J. Klerman, "The Long-Term Gains from GAIN: A Re-analysis of the Impacts of the California GAIN Program," Department of Economics, UCLA, unpublished manuscript (2001).
Hotz, J., G. Imbens, and J. Mortimer, "Predicting the Efficacy of Future Training Programs Using Past Experiences," forthcoming, *Journal of Econometrics* (2003).
Ichimura, H., and O. Linton, "Asymptotic Expansions for Some Semiparametric Program Evaluation Estimators," Institute for Fiscal Studies, cemmap working paper cwp04/01 (2001).
Ichimura, H., and C. Taber, "Direct Estimation of Policy Effects," Department of Economics, Northwestern University, unpublished manuscript (2000).
Imbens, G., "The Role of the Propensity Score in Estimating Dose-Response Functions," *Biometrika* 87:3 (2000), 706–710.
——— "Sensitivity to Exogeneity Assumptions in Program Evaluation," *American Economic Review Papers and Proceedings* (2003).
Imbens, G., and J. Angrist, "Identification and Estimation of Local Average Treatment Effects," *Econometrica* 61:2 (1994), 467–476.
Imbens, G., W. Newey, and G. Ridder, "Mean-Squared-Error Calculations for Average Treatment Effects," Department of Economics, UC Berkeley, unpublished manuscript (2003).
LaLonde, R. J., "Evaluating the Econometric Evaluations of Training Programs with Experimental Data," *American Economic Review* 76 (1986), 604–620.
Lechner, M., "Earnings and Employment Effects of Continuous Off-the-Job Training in East Germany after Unification," *Journal of Business and Economic Statistics* 17:1 (1999), 74–90.
Lechner, M., "Identification and Estimation of Causal Effects of Multiple Treatments under the Conditional Independence Assumption," in M. Lechner and F. Pfeiffer (Eds.), *Econometric Evaluations of Active Labor Market Policies in Europe* (Heidelberg: Physica, 2001).
——— "Program Heterogeneity and Propensity Score Matching: An Application to the Evaluation of Active Labor Market Policies," this REVIEW 84:2 (2002), 205–220.
Lee, D., "The Electoral Advantage of Incumbency and the Voter's Valuation of Political Experience: A Regression Discontinuity Analysis of Close Elections," Department of Economics, University of California, unpublished manuscript (2001).
Lehman, E., *Nonparametrics: Statistical Methods Based on Ranks* (San Francisco: Holden-Day, 1974).
Manski, C., "Nonparametric Bounds on Treatment Effects," *American Economic Review Papers and Proceedings* 80 (1990), 319–323.
Manski, C., G. Sandefur, S. McLanahan, and D. Powers, "Alternative Estimates of the Effect of Family Structure During Adolescence on

High School," *Journal of the American Statistical Association* 87:417 (1992), 25–37.

——— *Partial Identification of Probability Distributions* (New York: Springer-Verlag, 2003).

Neyman, J., "On the Application of Probability Theory to Agricultural Experiments. Essay on Principles. Section 9" (1923), translated (with discussion) in *Statistical Science* 5:4 (1990), 465–480.

Politis, D., and J. Romano, *Subsampling* (Springer-Verlag, 1999).

Porter, J., "Estimation in the Regression Discontinuity Model," Harvard University, unpublished manuscript (2003).

Quade, D., "Nonparametric Analysis of Covariance by Matching," *Biometrics* 38 (1982), 597–611.

Robins, J., and Y. Ritov, "Towards a Curse of Dimensionality Appropriate (CODA) Asymptotic Theory for Semi-parametric Models," *Statistics in Medicine* 16 (1997), 285–319.

Robins, J. M., and A. Rotnitzky, "Semiparametric Efficiency in Multivariate Regression Models with Missing Data," *Journal of the American Statistical Association* 90 (1995), 122–129.

Robins, J. M., Rotnitzky, A., Zhao, L.-P., "Analysis of Semiparametric Regression Models for Repeated Outcomes in the Presence of Missing Data," *Journal of the American Statistical Association* 90 (1995), 106–121.

Rosenbaum, P., "Conditional Permutation Tests and the Propensity Score in Observational Studies," *Journal of the American Statistical Association* 79 (1984a), 565–574.

——— "The Consequences of Adjustment for a Concomitant Variable That Has Been Affected by the Treatment," *Journal of the Royal Statistical Society, Series A* 147 (1984b), 656–666.

——— "The Role of a Second Control Group in an Observational Study" (with discussion), *Statistical Science* 2:3 (1987), 292–316.

——— "Optimal Matching in Observational Studies," *Journal of the American Statistical Association* 84 (1989), 1024–1032.

——— *Observational Studies* (New York: Springer-Verlag, 1995).

——— "Covariance Adjustment in Randomized Experiments and Observational Studies," *Statistical Science* 17:3 (2002), 286–304.

Rosenbaum, P., and D. Rubin, "The Central Role of the Propensity Score in Observational Studies for Causal Effects," *Biometrika* 70 (1983a), 41–55.

——— "Assessing the Sensitivity to an Unobserved Binary Covariate in an Observational Study with Binary Outcome," *Journal of the Royal Statistical Society, Series B* 45 (1983b), 212–218.

——— "Reducing the Bias in Observational Studies Using Subclassification on the Propensity Score," *Journal of the American Statistical Association* 79 (1984), 516–524.

——— "Constructing a Control Group Using Multivariate Matched Sampling Methods That Incorporate the Propensity Score," *American Statistician* 39 (1985), 33–38.

Rubin, D., "Matching to Remove Bias in Observational Studies," *Biometrics* 29 (1973a), 159–183.

——— "The Use of Matched Sampling and Regression Adjustments to Remove Bias in Observational Studies," *Biometrics* 29 (1973b), 185–203.

——— "Estimating Causal Effects of Treatments in Randomized and Non-randomized Studies," *Journal of Educational Psychology* 66 (1974), 688–701.

——— "Assignment to Treatment Group on the Basis of a Covariate," *Journal of Educational Statistics* 2:1 (1977), 1–26.

——— "Bayesian Inference for Causal Effects: The Role of Randomization," *Annals of Statistics* 6 (1978), 34–58.

——— "Using Multivariate Matched Sampling and Regression Adjustment to Control Bias in Observational Studies," *Journal of the American Statistical Association* 74 (1979), 318–328.

Rubin, D., and N. Thomas, "Affinely Invariant Matching Methods with Ellipsoidal Distributions," *Annals of Statistics* 20:2 (1992), 1079–1093.

Seifert, B., and T. Gasser, "Finite-Sample Variance of Local Polynomials: Analysis and Solutions," *Journal of the American Statistical Association* 91 (1996), 267–275.

——— "Data Adaptive Ridging in Local Polynomial Regression," *Journal of Computational and Graphical Statistics* 9:2 (2000), 338–360.

Shadish, W., T. Campbell, and D. Cook, *Experimental and Quasi-experimental Designs for Generalized Causal Inference* (Boston: Houghton Mifflin, 2002).

Sianesi, B., "psmatch: Propensity Score Matching in STATA," University College London and Institute for Fiscal Studies (2001).

Smith, J. A., and P. E. Todd, "Reconciling Conflicting Evidence on the Performance of Propensity-Score Matching Methods," *American Economic Review Papers and Proceedings* 91 (2001), 112–118.

——— "Does Matching Address LaLonde's Critique of Nonexperimental Estimators," forthcoming, *Journal of Econometrics* (2003).

Van der Klaauw, W., "A Regression-Discontinuity Evaluation of the Effect of Financial Aid Offers on College Enrollment," *International Economic Review* 43:4 (2002), 1249–1287.

Zhao, Z., "Using Matching to Estimate Treatment Effects: Data Requirements, Matching Metrics, and Monte Carlo Evidence," this REVIEW this issue (2004).